Statistical Optimization of Leakage Power Considering Process Variations using Dual-Vth and Sizing*

Ashish Srivastava Dennis Sylvester David Blaauw
University of Michigan, EECS Department, Ann Arbor, MI 48109
{ansrivas, dennis, blaauw}@eecs.umich.edu

Abstract

Increasing levels of process variability in sub-100nm CMOS design has become a critical concern for performance and power constraint designs. In this paper, we propose a new statistically aware Dual-Vt and sizing optimization that considers both the variability in performance and leakage of a design. While extensive work has been performed in the past on statistical analysis methods, circuit optimization is still largely performed using deterministic methods. We show in this paper that deterministic optimization quickly looses effectiveness for stringent performance and leakage constraints in designs with significant variability. We then propose a statistically aware dual-Vt and sizing algorithm where both delay constraints and sensitivity computations are performed in a statistical manner. We demonstrate that using this statistically aware optimization, leakage power can be reduced by 15-35% compared to traditional deterministic analysis. The improvements increase for strict delay constraints making statistical optimization especially important for high performance designs.

Categories and Subject Descriptors: B.6.3 Performance Analysis and Design Aids

General Terms: Algorithms, performance, reliability

Keywords: Leakage, variability, optimization

1. Introduction

Traditionally designers have used case files, or corner case models, to optimize and ascertain the performance of their designs. Best, worst, and nominal case models for the devices are developed and the design is required to meet specifications at all process corners. However, this approach can both significantly over- and underestimate the impact of the underlying variations on the design. Overestimation makes the specification, typically timing, harder to meet, leading to increased design time/effort and results in lost performance. On the other hand underestimation can lead to yield loss [1]. Furthermore, case files provide very limited information to the designers when they attempt to perform yield-based optimization and robustness analysis.

The increase in process variation with technology scaling has made worst-case analysis unacceptable [2]. Thus, statistical modeling of circuit performance is now imperative. Recently, various studies have been conducted to estimate the impact of variability on performance and yield. For example, [3,4] address the impact of process variation on the distribution of circuit and path delays. In [5,6,7], the authors develop statistical timing analysis tools to replace standard deterministic static timing analyzers whereas [8,27] develop approaches for the statistical estimation of leakage power considering within-die and across-die variations.

However, very little work has been done on using statistical approaches to perform circuit optimization. Previous work [9,10] uses joint probability density functions (PDFs) of the circuit performance metrics and poses the yield optimization problem as a maximization of a higher dimensional integral which are estimated using Monte Carlo simulations. However, these methods are difficult to apply in modern applications due to their high runtime and memory requirements with increases in statistical parameters.

Recent approaches to counter the impact of process variation have generally been limited to post-fabrication techniques. Forward and reverse body-biasing have been shown to improve yield and result in tighter distributions of circuit performance [11]. Reference [12] compares the approaches of adaptive body-bias and adaptive power supply to counter process variability. In [13], a simple circuit structure is used to automatically generate the ideal body-bias which is a function of process parameters and is ideal for a localized portion of the die. Alternatively, [14] proposes an optimization method to counter the effects of process variations. However this approach does not actually use statistical analysis but instead employs a heuristic to prevent a buildup of critical timing paths during the optimization.

Thus, we see that although a large amount of work is aimed towards countering the effects of process variations, there is only limited effort thus far in developing optimization approaches that consider these effects making intelligent decisions based on statistical information.

The tremendous impact of variability was demonstrated recently in [11], showing 20X variation in leakage power for a 1.3X variation in delay between fast and slow dies. Due to the inverse relationship between leakage power and gate delay, most of the fastest chips in a lot are found to have unacceptable leakage and vice-versa. In addition, low-V_{th} devices, which are used in now-common dual-V_{th} processes, exhibit increased sensitivity to variations [15] in their leakage power. Figure 1 compares the PDFs and cumulative distribution function

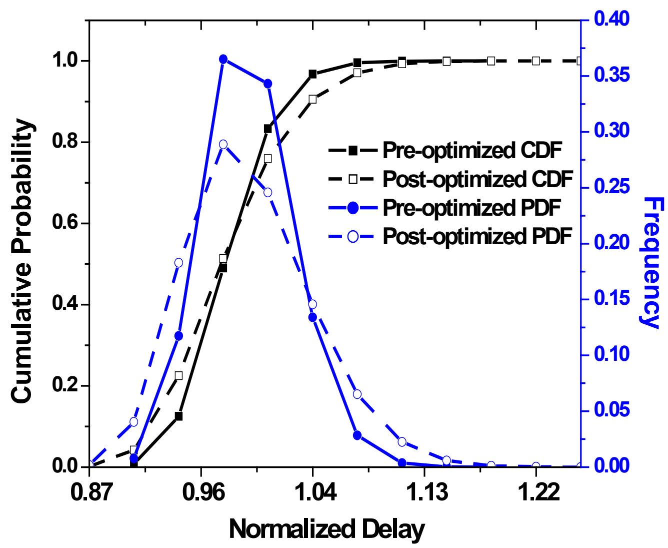


Figure 1. Impact of low-power optimization approaches on delay PDFs and CDFs.

*This work was supported in part by funding from NSF, SRC, GSRC, IBM and Intel.

(CDFs) of a pre- and post-optimized design. The pre-optimized design refers to a design optimally sized to meet a delay target with just one threshold voltage. The design is then optimized for leakage power using an additional threshold voltage [16] while nominal delay is constrained to remain identical. Note that the post-optimized PDF exhibits many more paths at the slow-end of the distribution which indicates a parametric yield loss. Based on this figure, we see that there is a pressing need to devise optimization approaches that make use of available statistical information to simultaneously improve yield and performance.

In this paper, we propose an approach to minimize leakage power using a dual-V_{th} approach coupled with sizing while considering the impact of process variation. All previous approaches for dual-V_{th} assignment [16-21] have neglected the impact of such variations, hence using these approaches in current technologies can adversely impact both yield and performance.

The remainder of the paper is organized as follows. Section 2 discusses the models and analysis methods for statistical timing and leakage power analysis. Section 3 presents the statistical enhancements made to a traditional dual-Vth and sizing algorithm. In Section 4, we present results and compare our algorithm to a traditional deterministic optimization. We summarize and conclude in Section 5.

2. Preliminaries

The traditional approach of case file-based optimization has been able to capture die-to-die variations, but results in very pessimistic results when used to model within-die variations [5]. In this work we only consider variations in the within-die component of gate length, usually considered to be the dominant variation source in most circuits [2]. Since gate length also strongly impacts V_{th} we also implicitly model V_{th} uncertainty. Though our present work only considers variations in gate length, the approach in general can very easily be extended to multiple parameters varying simultaneously. To capture the impact of this variation, a standard cell library is characterized for delay and leakage power variation with varying gate length. All transistors within a gate are assumed to be perfectly correlated and variation is assumed to be independent across gates. Assuming a total gate length variation of 15%, the within-die component is estimated by dividing this total variation budget equally into within-die and across-die variation components [3].

2.1 Statistical Delay Estimation

A quadratic model is used to express the dependency of delay on gate length (L_{gate}) as shown in Equation (1).

$$Delay = f(L_{gate}) = a_0 + a_1 L_{gate} + a_2 L_{gate}^2 \quad (1)$$

Gates are characterized at seven different capacitive gate loads and seven input transition times to generate a table-lookup for each of the fitting parameters used in the quadratic model. The mean and variance of gate delays can then be expressed in terms of the higher order moments of the gate length as

$$Mean = a_0 + a_1 E[L_{gate}] + a_2 E[L_{gate}^2] \quad (2)$$

$$Var = a_2^2 E[L_{gate}^4] + 2a_1 a_2 E[L_{gate}^3] - \left(a_2^2 E[L_{gate}^2]\right)^2 - 2a_1 a_2 E[L_{gate}][L_{gate}^2]$$

Since the gate lengths are assumed to be Gaussian, the higher order moments of the gate length can be obtained using the following relations,

$$E[(Delay - \mu)^{2k}] = (2k)!\sigma^{2k} / (k!2^k) \quad (3)$$

$$E[(Delay - \mu)^{2k+1}] = 0$$

where μ and σ are the mean and standard deviation of the gate length and k is any positive integer. The mean and variance of the gate lengths are then used by a statistical STA (SSTA) [5] engine to predict PDFs of delay at each node in the circuit. The SSTA tool assumes the gate delays are normally distributed. This assumption is reasonable since the quadratic dependence of delay on the gate length is generally very weak; we primarily use a quadratic fit to enable accurate estimations of the mean and variance of the gate delay. The SSTA engine employs a discrete version of these PDFs to enable efficient computation of delay PDFs within the circuit. These delay PDFs can then be used to determine any confidence point of the delay.

We observe that the gates using low-V_{th} transistors show smaller variation in delay. Assuming that the variation in threshold voltage of the low-V_{th} gates is not larger than the high-V_{th} gates. The smaller impact on delay can be understood by using the alpha-power law model [22]. This model can be used to express the sensitivity of gate delay with respect to V_{th} as

$$S_{Delay}^{V_{th}} = \frac{V_{th}}{Delay} \frac{\partial Delay}{\partial V_{th}} \alpha \frac{V_{th}}{(Vdd - V_{th})} \quad (4)$$

which shows that the impact of V_{th} variation on delay reduces with an increase in the difference between the supply and threshold voltage. Therefore, a larger variation in delay for the high V_{th} gates is expected [4]. It is interesting to note that though the low V_{th} gates are less susceptible to delay variations they are highly susceptible to leakage power variations [15].

The above statistical delay modeling approach can easily be extended to multiple sources of variation assuming the sources are independent. This is achieved by first expressing the delay as a function of the required parameters (as in Equation (1)). Based on parameter independence, we can then develop simple expressions for the mean and variance of the gate delays in terms of the moments of the varying parameters.

2.2 Statistical Leakage Power Estimation

We capture the dependence of leakage power on gate length using an exponential decay model:

$$Power_{Leak} = g(L_{gate}) = a_0 \exp\left(\frac{-L_{gate}}{a_1}\right) = \exp\left(\frac{-L_{gate}}{a_1} + \ln(a_0)\right) \quad (5)$$

This allows us to express the leakage power PDF of each gate as a lognormal where the corresponding Gaussian is a linear transformation of L_{gate}. The PDF of total circuit leakage can then be expressed as a sum of independent lognormals, which can be well approximated by another lognormal using Wilkinson's method [23]. This approximation is valid even in the presence of a correlated component of process variation.

$$S = X_1 + X_2 + ... = e^{Y_1} + e^{Y_2} + ... = e^Z \quad (6)$$

In Wilkinson's method the mean and the variance of S are obtained by matching the first two moments of S and (X_1+X_2+…). The mean and variance of S are then used to calculate the parameters of the lognormal, defined to be the mean and variance of the corresponding gaussian [23]. The PDF of the block leakage current is then given by

$$f(i_{leak}) = \left(\frac{1}{x\sqrt{2\pi}\beta}\right) \exp\left(\frac{-(\ln(i_{leak}) - \alpha)^2}{2\beta^2}\right) \quad (7)$$

where α and β are the parameters of the lognormal distribution. For a lognormal distribution the percent point function is defined as:

$$\Pr[X \le f(\theta)] = \theta \quad (8)$$

This can be expressed in terms of the percent point function of the normal distribution Φ^{-1} as [24]:

$$f(\theta) = \exp\left(\alpha + \beta\Phi^{-1}(\theta)\right). \quad (9)$$

which can then be used to estimate the confidence points of a lognormal distribution.

This approach is again readily extended to additional sources of variation. In particular, the variation in V_{th} can be expressed as an exponential multiplicative term in Equation (5). The leakage power then becomes an exponential of a sum of two normal distributions which has the same form as Equation (7).

3. Statistical Dual-V_{th} Assignment with Sizing

The statistical dual-V_{th} and sizing problem is expressed as an assignment problem that seeks to find an optimal assignment of threshold voltages (from a set of two thresholds) and drive strengths (from a set of drive strengths available in a standard cell library) for each of the gates in a given circuit network. The objective is to minimize the leakage power *measured at a high percentile point of its PDF* while maintaining a timing constraint imposed on the circuit. The timing constraint is also expressed as a *delay target for a high percentile point* of the circuit delay. These timing and power constraints can be determined based on desired yield estimates, such as 95% or 99%. This is opposed to an ideal approach where the yield of the circuit is expressed in terms of a joint PDF of the delay and leakage power and the yield is then optimized at a high percentile point. Our formulation in this work serves to simplify the problem and allows traditional iterative optimization approaches to be easily adapted to statistical optimization.

First we outline a traditional deterministic dual-V_{th} and sizing approach [16] that uses corner case files and then introduce two approaches to include the effects of process variation in low-power optimization.

3.1 Deterministic approach

The initial design, using the lower V_{th} exclusively, is first sized to meet the timing constraint using a TILOS-like optimizer [25]. A sensitivity measure is defined as

$$S_{gate}^{swap} = \frac{|\Delta P|}{|\Delta D|} Slack_{gate} \tag{10}$$

and evaluated for all low V_{th} gates in the circuit. ΔP and ΔD in Equation (10) are the changes in power dissipation and delay of the gate when the low-V_{th} gate is swapped with a high-V_{th} gate (of same size and functionality). The gate with the maximum sensitivity (e.g., G1) is then swapped with a high-V_{th} version of the gate. If the circuit now fails to meet timing a new sensitivity measure is defined as

$$S_{gate}^{up-size} = \frac{1}{\Delta P} \sum_{arcs} \frac{\Delta D}{Slack_{arc} - S_{\min} + K} \tag{11}$$

where S_{min} is the worst slack seen in the circuit and K is a small positive quantity to maintain stability. Equation (11) is then evaluated for all gates in the circuit. This form of the sensitivity metric places a higher weighting to gates lying on the critical paths of the circuit. The arcs over which the summation is taken represent the falling and rising arcs associated with each of the inputs of the gate. Thus, for a 3-input NAND gate the sensitivity measure will be obtained by summing over all six possible arcs.

deterministic dual-V_{th}

0: $Power_0$=calculate power
1: Calculate sensitivity (S^{swap}) of low V_{th} gates
2: Set gate with maximum S^{swap} to high V_{th}
3: check timing
4: if circuit meets timing goto STEP0
4: Calculate sensitivity ($S^{up-size}$) for all gates
5: up-size gate with maximum $S^{up-size}$
6: if (power > $Power_0$) undo moves and goto STEP0
7: goto STEP4

when the gate is up-sized to the next available size in the library. The gate with the maximum sensitivity is then up-sized and the process is repeated until either the circuit meets timing or the power dissipation increases relative to its level prior to gate G1 being set to high-V_{th}. In the latter event gate G1 is set back to high-V_{th} and is flagged to prevent the gate from again being considered for high V_{th} assignment again later in the process.

3.2 Statistical approach

We propose two major enhancements to the above deterministic approach that use available statistical information to improve the overall optimization. In the first improvement the timing check in STEP3 of the deterministic dual-V_{th} approach is performed using statistical timing analysis. The required percentile point on the delay PDF, used to specify the constraint, is now obtained from the PDFs generated by the SSTA engine rather than a corner model case file.

A deterministic timing analyzer is used to determine the input slope at each of the gates, which is then used along with the output capacitance as indices in the look-up table for the fitting parameters (in Equations (1) and (5)). The mean and variance are estimated using Equations (2)-(3) and are then passed onto the SSTA engine to evaluate the PDF of the arrival and required times at each circuit node. Note that while performing the statistical timing analysis, additional dummy source and sink nodes are added to the circuit, hence the delay constraint needs to be checked at just one point within the network [5]. Using statistical delay analysis reduces the pessimism in timing since all gates cannot be expected to be simultaneously operating at their worst-case corners, an assumption that is made when performing a corner-based worst-case analysis. We show in Section 4 that optimizing a circuit to meet a delay constraint using worst-case analysis results in a substantial loss in circuit performance optimality. The situation is worsened for leakage power optimization because of the exponential dependence of leakage power on threshold voltage.

The second enhancement uses the statistical information in the fitting functions of delay and power to guide the optimization by replacing the sensitivities evaluated in STEP1 and STEP4 with *statistical sensitivities*. These statistical sensitivities are then evaluated at a confidence point on the PDF of the sensitivity. Since generating PDFs of the sensitivity metrics themselves is fairly complicated and computationally intensive, we estimate the statistical sensitivities by evaluating the mean and standard deviation of these PDFs (i.e., we only concern ourselves with the first and second central moments of the sensitivity PDFs and not their entire shape). Also, the dependence of slack on gate length of the devices is not straightforward and we make the assumption that the slack is independent of gate length while calculating the moments of the sensitivities. The sensitivities in Equations (10)-(11) can now be expressed as a product of two independent random variables X and Y where X is dependent on L_{gate} and Y is not. Thus X corresponds to the ratio of the change in power and change in delay, and Y corresponds to the slack dependent terms in Equations (10)-(11). Given two independent random variables X and Y, the expectation of their product is the same as the product of their expectation. Using this fact, we can estimate the mean and standard deviation of the sensitivities using the independence assumption made above and using the following relations:

$$\begin{aligned} E(XY) &= E(X)E(Y) \\ Var(XY) &= E((XY - E(XY))^2) = E(X^2)E(Y^2) - (E(XY))^2 \end{aligned} \tag{12}$$

where E(X) is the expected value of X alone and E(Y) is the expected value of Y alone.

The mean and variance of the terms involving L_{gate} (X in Equation 12) are expressed as a function '*f*' of L_{gate} alone, using the delay and power models (Equations (1) and (5)). The expected value is then written as

$$E(f(L_{gate})) = \int_{-\infty}^{\infty} f(L_{gate})p(L_{gate})dL_{gate} \tag{13}$$

where $p(L_{gate})$ is the PDF of the gate length. Applying Taylor series theorem to the above expression we can re-write it as follows

$$E(f(L_{gate})) = \int_{-\infty}^{\infty} \left(f(\mu) + f'(\mu)(L_{gate} - \mu) + ... + f^n(\mu)\frac{(L_{gate} - \mu)^n}{n!} \right) p(L_{gate})dL_{gate} \tag{14}$$

where μ is the mean of L_{gate}. This gives

$$E(f(L_{gate})) = f(\mu) + \frac{f''(\mu)\eta_2}{2!} + ... + f^{2n}(\mu)\frac{\eta_{2n}}{(2n)!} \tag{15}$$

where η_i is the i'th central moment of L_{gate} and the odd central moments of L_{gate} are set to zero. This approximation can be used to obtain the mean and variance of the sensitivities. For our analysis, we found that a fourth-order approximation of $f(L_{gate})$ was sufficient for good accuracy.

The moments of the slack dependent terms (Y in Equation 12) are estimated by using the slack PDFs obtained from SSTA. The statistical sensitivities are now redefined by evaluating at a certain number '*n*' standard deviations away from the mean. Since the shape of the sensitivity PDFs is not known, *'n'* is not known even if a known confidence point is desired. Later in Section 4 we look at the impact of *'n'* on the optimization results.

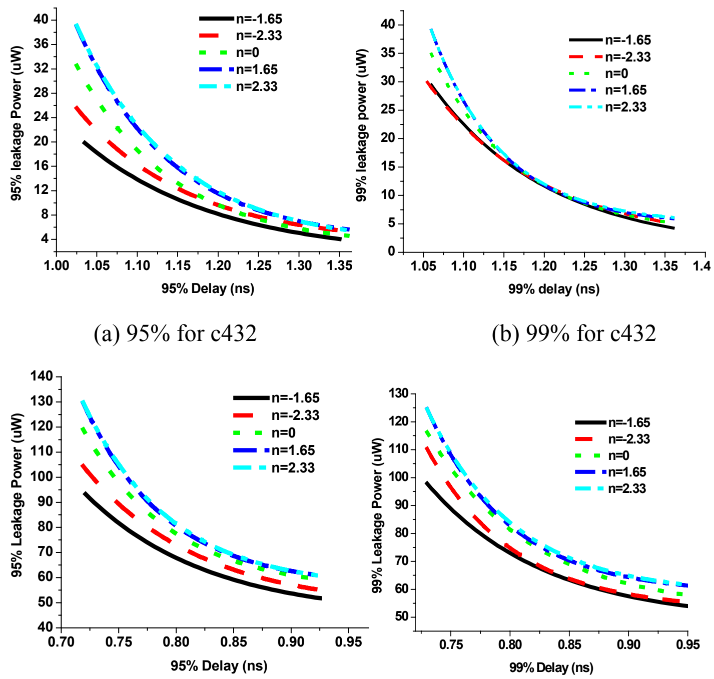
We note that the approach can be easily extended to multiple delay constraints, where a set of percentile points on the delay PDF can be constrained to be less than some desired value. As an example, this flexibility is well suited to microprocessor designs where we can simultaneously constrain the 95^{th} and 99^{th} percentile delay to concurrently target different yields for different performance bins.

The worst-case time complexity of the algorithm can be expected to be $O(n^3)$ since the SSTA engine has a linear time complexity [5] and in the worst-case we could up-size the entire circuit each time we set a gate to high V_{th}. This would happen when we maximally size-up the circuit each time we set a gate to high V_{th} yet still fail to meet timing (this also requires the total power not to surpass the original circuit through all up-sizing moves). Note that in the worst-case the $O(n^3)$ complexity results because the total number of up-sizing moves (and reversed up-sizing moves) is $O(n^2)$ since every gate is up-sized to the maximum size available in the library whenever a gate is set to high V_{th}, and all the moves are then reversed. If the total number of up-sizing moves that are reversed is assumed to be linearly proportional to the number of gates in the circuit, the overall complexity of the algorithm reduces to $O(n^2)$, since the total number of up-sizing or cell-swapping moves now become linearly proportional to the number of gates in the circuit.

4. Results

The benchmark circuits are synthesized using an industrial 0.13μm standard cell library with a V_{dd} of 1.2V and a high and a low V_{th} of 0.23V and 0.12V respectively. For the delay constraints, we consider two different cases where the delay is constrained at the 95^{th} or 99^{th} percentile. Leakage power is optimized at the same percentile point used to express the delay constraint.

To make a fair comparison of the statistical and deterministic approaches, the best and worst-case corner models for the gates are developed for the same percentile point at which the delay constraint was specified for a particular experiment (95^{th} or 99^{th}).



(a) 95% for c432 (b) 99% for c432

(c) 95% for c499 (d) 99% for c499

Figure 2. Impact of '*n*' on statistical optimization

Figure 2 shows the impact of evaluating the sensitivities at different points along their distribution (relative to the mean) on the final optimization results for two ISCAS'85 [26] benchmark circuits. The sensitivities are evaluated at a fixed number of standard deviations away from the mean which is represented as *'n'* (Section 3.2). The curves are obtained through multiple runs of the algorithm. Each time the algorithm is run, the delay constraint is progressively tightened to obtain a complete power-delay curve. For both 95^{th} and 99^{th} percentile delay constraints, we observe that considering $n = -1.63$ (corresponding to the 5^{th} percentile point on a Gaussian) leads to the best power-delay curve characteristics. For the 99^{th} percentile case we observe that both n=-1.63 and n=-2.33, which corresponds to the 1st percentile point in a Gaussian perform very similarly. The significant improvement over the cases where a high percentile point of the sensitivities is used to select the gate to be swapped/up-sized can be understood by noting that a low percentile point on the sensitivity point gives a high confidence that the sensitivity value is at least as large as the value at the decision-making point.

Figure 3 compares three different optimization approaches outlined in Section 3. In particular we sub-divide the statistical optimization approach of Section 3.2 into two stages – 1) "with statistical constraints" which relies on SSTA but does not include statistical sensitivities, and 2) "with statistical sensitivities" which includes both improvements described above. The 95% delay and 95% power are estimated using the statistical estimation techniques discussed in Section 2 for all curves except "delay using corner models". It is interesting to note that the incorporation of statistical sensitivities provides an additional reduction of ~40% in leakage power at the tightest delay constraint compared to the case where we only use the SSTA engine to enforce the delay constraint. This indicates that although the use of a statistical timing analysis framework is clearly important, statistically modeling the power and delay impact of change in V_{th} is equally critical. Additionally, the optimization based on corner models (using the traditional approach of Section 3.1) is not able to meet very tight constraints on the 95^{th} percentile of the delay that are met by optimizations that employ an SSTA engine due to the pessimism of the corner model approach. The last curve ("delay using corner models") plots the results for the optimization using corner models where the delay is calculated using worst-case models.

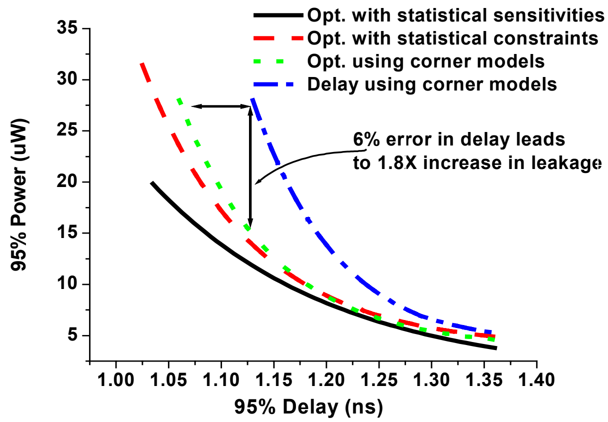


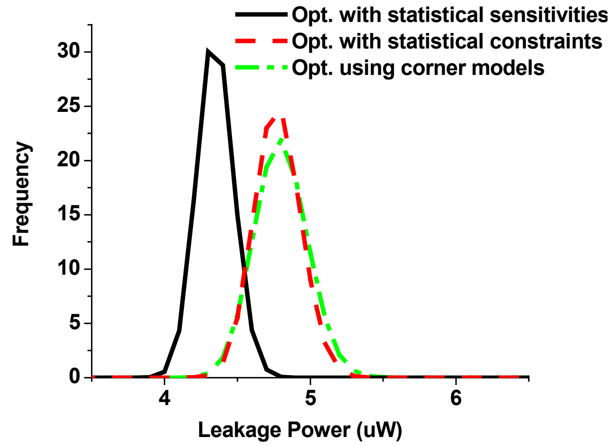
Figure 3. Power-delay curves for the three optimization techniques

Table 1. Power savings for the statistical approaches compared to a corner-model based approach.

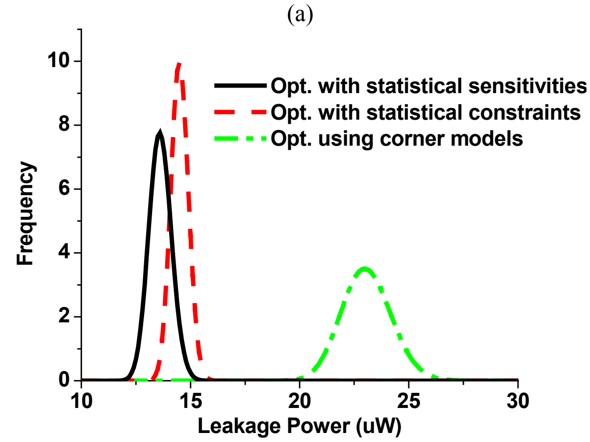
Circuit	Power (95%)		Power (99%)		Gate Count	RunTime (min)
	OPT2	OPT3	OPT2	OPT3		
c432	16.3%	39.3%	18.0%	35.7%	165	1
c499	4.3%	30.7%	23.9%	30.3%	519	13
c880	9.5%	13.2%	8.0%	50.0%	390	8
c1908	12.7%	23.6%	11.5%	35.5%	432	10
c2670	5.3%	20.3%	36.8%	45.6%	965	20
c3540	35.3%	43.5%	7.3%	15.3%	962	19
c5315	17.8%	34.3%	31.1%	39.6%	1750	68
c6288	15.0%	26.5%	22.4%	36.1%	2502	115
Average	14.5%	28.9%	19.9%	36.0%		

The curve shows that if statistical information is not provided to the designer a small overestimation in the delay can result in large performance improvements being left on the table, since designs are generally optimized within a strict delay constraint. Also high-performance circuits generally operate in a steep region of the power-delay curve and a small overestimation in delay can be expected to result in a large loss in the achievable improvements of the performance parameter being optimized. It can be seen that the different optimization cases also tend to converge as the delay constraint is relaxed. This can be understood by noting that as the delay constraints is relaxed a larger fraction of gates are assigned to high V_{th} and hence the final state becomes increasingly independent of the order in which the gates were assigned to high V_{th}.

Table 1 summarizes the improvement in leakage power for the ISCAS'85 benchmark circuits for the statistical optimization approaches described in Section 3.2 compared to a deterministic approach. OPT2 and OPT3 refer to the optimization with statistical timing constraints alone and with both statistical timing constraints and sensitivities, respectively. The results are shown for the best delay constraint met using the corner models, thus the results in Table 1 for the 95th and 99th percentile cases correspond to different delay constraints. Average reductions in leakage power of approximately 14% and 29% can be achieved by using OPT2 and OPT3, respectively, for the 95th percentile case compared to a traditional deterministic approach. A larger average improvement of approximately 20% and 36% can be obtained for the 99th percentile case. These delay points correspond to the high frequency bin and are most affected by leakage power dissipation. The last columns of the table list the size of the circuits and the runtime for the algorithm on an Intel 2.8GHz Xeon processor with 3GB of RAM. We observe that runtime follows the quadratic complexity predicted in Section 3.2. Figure 4 compares the PDF of leakage power for the three optimization approaches for both loose and tight delay constraints**.** These power curves are all taken with identical 95% delays, or identical performance. For loose delay constraints (Figure 4a) all three optimization approaches result in fairly similar PDFs for leakage power. This again reflects the fact that the different optimization approaches behave very similarly for loose delay constraints.



(a)



(b)

Figure 4. PDFs of leakage power for (a) loose delay constraints (b) tight delay constraints.

The tighter constraints clearly separate the leakage power PDFs of the statistical and deterministic approaches. It is interesting to note that although statistical sensitivities lead to a smaller 95th percentile leakage power as compared to the other approaches, the variance is marginally larger when compared to the optimization using only statistical constraints. We emphasize that Figure 4b corresponds to the highest performance parts being manufactured and using statistical optimization leads to not only a much smaller average leakage power but also reduces the spread of the distribution considerably which significantly impacts the yield.

5. Conclusions

We present an approach to use statistical information to make effective decisions while performing low-power optimization. The simplicity of our approach makes it amenable to inclusion within already existing sensitivity-based optimization approaches. We have demonstrated this by implementing the new techniques within an already existing dual-V_{th} and sizing algorithm and shown the advantages offered by statistical optimization in comparison with traditional corner model based optimizations. The results obtained show that a reduction of

~15-35% in leakage power can be obtained on average for the high frequency bins of the design. We also show that statistical optimization leads to much tighter distributions of power, which is ideal both from performance and yield perspectives.

Acknowledgement

The authors would like to acknowledge the assistance provided by Aseem Agarwal with the statistical timing analyzer.

References

[1] S. Duvall, "Statistical Circuit Modeling and Optimization," *Workshop on statistical metrol*ogy, pp.56-63, 2000.
[2] S. Nassif, "Delay variability: sources, impacts and trends," *Proc. ISSCC,* pp. 368-369, 2000.
[3] K. Bowman *et al*., "Impact of die-to-die and within-die Parameter Fluctuations on the Maximum Clock Frequency Distribution for Gigascale Integartion", *IEEE J. Solid-State Circuits,* pp.183-190, Feb. 2002.
[4] M. Eisele, *et al.,* "The impact of intra-die device parameter variations on path delays and on the design for yield of low voltage digital circuits," *IEEE Transactions on VLSI Systems*, Dec. 1997.
[5] A. Agarwal, *et al.,* "Computation and refinement of statistical bounds on circuit delay," *Proc. DAC,* pp.348-353, 2003.
[6] J. Jess et al., "Statistical timing for parametric yield prediction of digital integrated circuits," *Proc. DAC,* pp.932-937, 2003.
[7] M. Orshansky and K. Keutzer, "A general probabilistic framework for worst case timing analysis," *Proc. DAC*, pp.556-561, 2002.
[8]. R. Rao, *et al.,* "Statistical estimation of leakage current considering inter and intra-die process variations", *Proc. ISLPED*, pp.84-49, 2003.
[9] P. Feldmann and S. Director, "Accurate and efficient evaluation of Circuit yield and yield gradients", *Proc. ICCAD*, pp.120-123, 1990.
[10] S. Director, *et al.,* Optimization of Parametric yield: A tutorial," *Proc. CICC,* 1992.
[11] S. Borkar, *et al.,* "Parameter variation and Impact on Circuits and Microarchitecture," *Proc. DAC,* pp.338-342, 2003.
[12] S. Naffziger and T. Chen, "Comparison of adaptive body bias and adaptive supply voltage for improving delay and leakage under the presence of process variation," *IEEE Tran. VLSI Systems*, pp.888-899, Oct 03.
[13] C. Neau and K. Roy, "Optimal body bias selection for leakage improvements and process compensation over different technology generations," *Proc. ISLPED,* pp.116-121, 2003.
[14] X. Bai et al., "Uncertainty aware circuit Optimization," *Proc. DAC,* pp.58-63, 2002.
[15] Ruchir Puri, personal communication.
[16] S. Sirichotiyakul, *et al.,* "Stand-by Power Minimization through Simultaneous Threshold Voltage Selection and Circuit Sizing," *Proc. DAC*, pp. 436-441, 1999.
[17] P. Pant, R. Roy, and A. Chatterjee. "Dual-threshold voltage assignment with transistor sizing for low power CMOS circuits," *IEEE Transactions on VLSI Systems,* pp.390-394, 2001.
[18] L. Wei, K. Roy, and C. Koh, "Power minimization by simultaneous dual-Vth assignment and gate sizing," *Proc. CICC,* pp.413-416, 2000.
[19] T. Karnik, *et al*., "Total power optimization by simultaneous dual-Vth allocation and device sizing in high performance microprocessors," *Proc. DAC,* pp.486-491, 2002.
[20] Qi Wang and S. Vrudhula, "Static power optimization of deep submicron CMOS circuits for dual V_T technology," *Proc. ICCAD,* pp.490-496, 1998.
[21] M. Ketkar and S. S. Sapatnekar, "Standby Power Optimization via Transistor Sizing and Dual Threshold Voltage Assignment," *Proc .ICCAD*, pp.375-378, 2002.
[22] T. Sakurai and A.R. Newton, "Alpha-power law MOSFET model and its application to CMOS inverter delay and other formulas," *IEEE J. Solid-State Circuits,* pp.584-593, April 1990.
[23] S.C. Schwartz and Y.S. Yeh, "On the distribution function and moments of power sums with lognormal components," *Bell Systems Technical Journal*, vol.61, pp.1441-1462, Sept. 1982.
[24] http://www.itl.nist.gov/div898/handbook/eda/eda.htm
[25] J. Fishburn and A. Dunlop, "TILOS: a posynomial programming approach to transistor sizing", *Proc. ICCAD,* pp.326-328, 1985.
[26] F. Brglez and H. Fujiwara. "A neutral netlist of 10 combinatorial benchmark circuits," *Proc. ISCAS,* 1985, pp.695-698.
[27] S. Narendra, *et al.,* "Full-chip sub-threshold leakage power prediction model for sub-0.18 μm CMOS," *Proc. ISLPED*, pp.19-23, 2002.