Robust Gate Sizing via Mean Excess Delay Minimization

Jason Cong[1,2], John Lee[3] and Lieven Vandenberghe[3]
[1]Department of Computer Science
[2]California NanoSystems Institute
[3]Department of Electrical Engineering
University of California at Los Angeles
cong@cs.ucla.edu, {lee, vandenbe}@ee.ucla.edu

ABSTRACT

We introduce mean excess delay as a statistical measure of circuit delay in the presence of parameter variations. The β-mean excess delay is defined as the expected delay of the circuits that exceed the β-quantile of the delay, so it is always an upper bound on the β-quantile. However, in contrast to the β-quantile, it preserves the convexity properties of the underlying delay distribution. We apply the β-mean excess delay to the circuit sizing problem, and use it to minimize the delay quantile over the gate sizes. We use the Analytic Centering Cutting Plane Method to perform the minimization and apply this sizing to the ISCAS '85 benchmarks. Depending on the structure of the circuit, it can make significant improvements on the 95%-quantile.

Categories and Subject Descriptors

G.3 [**Probability and Statistics**]: Probabilistic algorithms; B.8 [**Performance and Reliability**]: General

General Terms

Algorithms, Design

Keywords

robust gate sizing, process variation, geometric programming, conditional value-at-risk

1. INTRODUCTION

As transistors become smaller, the increasing effect of process variations may cause many circuits to fail [16]. The random variations in the gate lengths, oxide thicknesses and doping will increase the variations in the delay to a size too large to be ignored. In this context it becomes necessary to make designs with robustness in mind.

To increase robustness of the design, the gate sizes, threshold voltages and other circuit parameters can be strategically assigned to improve the distribution of the circuit delay, subject to power and area constraints. After the main design is completed, redundancy and regularity can also be added to improve the yield of the circuit [7].

In this paper we use circuit sizing to improve the timing robustness of a design. This is not a new area, and several approaches to this problem already exist. In [6] the *bin-yield loss function*, defined as the expected loss or penalty for circuits that exceed a given delay threshold, is minimized using stochastic optimization. Other groups have used slack redistribution to reallocate the statistical slack of each of the paths [15]. The most popular method in literature estimates a worst-case scenario for each gate, and then sizes the circuit according to this estimate [9, 14, 11]. These methods add a "padding" that is proportional to the standard deviation of the gate delay. For the sake of simplicity, these methods will be referred to as "padded delay methods." The deficiency of the padded delay methods is that it uses a conservative estimate and it cannot distinguish between correlated and uncorrelated variations.

In this paper we propose the *Mean Excess Delay* (MED) as a statistical measure of the delay in the presence of parameter variations. This measure is used in the finance industry to minimize the risk associated with the value of an investment portfolio [13]. In the context of circuits, we show that MED is a convex function of the logarithm of the gate sizes, and therefore well-suited for minimization. We also discuss a numerical algorithm for mean excess delay gate sizing, and present some encouraging numerical results.

In summary there are two main contributions in the paper:

1. The introduction of the Mean Excess Delay as a statistical measure of circuit delay. The mean excess delay preserves the convexity of the underlying delay model, and is therefore well-suited for minimization.

2. Numerical results that compare the padded delay method with the mean-excess delay method for gate sizing.

The remainder of the paper is organized as follows. Section II gives a background on the circuit sizing problem in its nominal and statistical forms. In Section III we introduce the Mean Excess Delay function, and explain its mathematical properties. Section IV briefly outlines the minimization algorithm we use. Results are shown in Section V.

2. CIRCUIT SIZING WITH VARIATIONS

In the circuit sizing problem, the sizes of each gate are selected to minimize the delay of a circuit with constraints on the power and area.

2.1 The Nominal Case

In the nominal case, the variations in delay are ignored, resulting in the following problem

$$\begin{array}{ll} \text{minimize} & T^{\text{nom}}(x) \\ \text{subject to} & A(x) \le A_{\max} \\ & P(x) \le P_{\max} \\ & x \ge 0. \end{array} \tag{1}$$

Here, the optimization variable x is a vector of the log-gate sizes (or more accurately, the log of the normalized gate scaling factors). The functions $T^{\text{nom}}(x)$, $A(x)$ and $P(x)$ represent the nominal delay, area and power of the circuit as a function of the log of the gate sizes x. We assume the functions $A(x)$, $P(x)$ and $T^{\text{nom}}(x)$ are posynomials in convex form [3]. This problem has been studied extensively, and can be solved efficiently via geometric programming [5, 8]. For a good tutorial on circuit optimization via geometric programming, see [2].

2.2 The Statistical Delay

The effects of the process variations on the delay are often modeled as follows [6]:

$$d_k(x,v) = (1 + v_k)\, d_k^{\text{nom}}(x). \tag{2}$$

In the above, $d_k^{\text{nom}}(x)$ is the nominal delay, $d_k(x,v)$ is the gate delay with the process variations, and v is a vector of zero-mean random variables with the same dimension as x (v_k denotes the k^{th} element of the vector v). The random variables v_i are not restricted to be independent or Gaussian, and may have correlations that are "gate by gate", "die-to-die", or by location. This model preserves convexity of $d_k^{\text{nom}}(x)$. That is, if $d_k^{\text{nom}}(x)$ is convex, then $d_k(x,v)$ will be convex as well for fixed v.

Note that the distribution and the correlations of the random variable are unrestricted, making this model general enough to handle a large class of variations. For example, correlations between the gate length or doping can be included in the random variable v_i. The only restriction is that the effect on the delay is multiplicative, and the standard deviation of the variations is independent of the size of the gate. This last restriction is discussed below.

A modification of this model makes the variance of the random variables v_i a function of the corresponding gate size:

$$d_k(x,v) = (1 + v_k e^{-\alpha x_k})\, d_k^{\text{nom}}(x) \tag{3}$$

where v_k is a random variable that is scaled by the function $e^{-\alpha x_k}$. Here α is a modeling parameter that is 0.5 in Pelgrom's model [12] or 0.3 according to recent work in [18]. The intuition behind this model is that the saturation current I_{sat} is better controlled as gate is made larger. Note that if $v_k \ge 0$ then this $d_k(x,v)$ is convex for fixed v. However, if $v_k < 0$, the expression is not convex, and does not have a posynomial representation.

To distinguish between the two models, we will call model (2) the size independent variation model, and model (3) the size dependent variation model. The net effect of both models above is to make the total delay of the circuit a function of the gate-scaling factors *and* the random variables. This turns the total delay itself into a random variable.

2.3 The Statistical Sizing Problem

To convert the nominal problem (1) into one that minimizes the statistical delay, we need to decide what it means to minimize a random variable. One obvious definition is to focus on the β-quantile $q_\beta(x)$, defined as:

$$q_\beta(x) = \inf\{t \mid \mathbf{P}(T(x,v) \le t) \ge \beta\}.$$

In the context of circuits, the β-quantile is the delay specification that will give a yield of β-percent.

Reformulating the deterministic problem with a β-percent yield gives the following problem

$$\begin{array}{ll} \text{minimize} & q_\beta(x) \\ \text{subject to} & A(x) \le A_{\max} \\ & P(x) \le P_{\max} \\ & x \ge 0. \end{array} \tag{4}$$

This problem, however, is difficult to solve. First, the cost function q_β is difficult to evaluate, because in practice there is no closed-form expression for the distribution of $T(x,v)$, although there are many ways of approximating it [17, 4, 19, 20]. A further difficulty is that the problem (4) is generally not convex, even when the nominal problem (1) is convex.

3. THE MEAN EXCESS DELAY

The β-Mean Excess Delay (β-MED) is a measure that is closely related to the β-quantile. For continuous distributions, it is the expected value of the tail of the delay, when the tail is measured past the β-quantile.

$$m_\beta(x) = \mathbf{E}\left[T(x,v) \,|\, T(x,v) \ge q_\beta(x)\right]. \tag{5}$$

A graphical interpretation of the mean excess delay is shown in Fig. 1. In this graph, we plot the probability density function (pdf) of $T(x,v)$. $q_{0.95}(x)$ is the 95% quantile, and the shaded area to the right is the tail of $T(x,v)$. The center of mass of this shaded region is the mean excess delay of the distribution, $m_{0.95}(x)$.

This measure is used in the finance industry to manage risk. In this context, $T(x,v)$ is the loss of a portfolio, q_β is the value-at-risk, and m_β is the conditional value-at-risk [13].

3.1 Properties of the Mean-Excess Delay

By definition, $m_\beta(x) \ge q_\beta(x)$, *i.e.*, the β-mean excess delay is an upper bound on the β-quantile. Thus, minimizing mean excess delay will indirectly minimize the β-quantile, and it makes sense to consider the problem

$$\begin{array}{ll} \text{minimize} & m_\beta(x) \\ \text{subject to} & A(x) \le A_{\max} \\ & P(x) \le P_{\max} \\ & x \ge 0. \end{array} \tag{6}$$

The important feature of this measure is that it can be minimized via convex optimization, if $T(x,v)$ is convex in x for fixed v [13].

To show this, we define the function

$$g_\beta(x,t) = t + \frac{1}{1-\beta}\mathbf{E}_v[T(x,v) - t]^+ \tag{7}$$

where

$$[u]^+ = \begin{cases} u & \text{if } u \ge 0 \\ 0 & \text{otherwise} \end{cases}$$

and $\mathbf{E}_v$ is expectation with respect to v, *i.e.*, if v is a continuous random vector in $\mathbb{R}^n$ with pdf $p(v)$,

$$\mathbf{E}_v[T(x,v)-t]^+ = \int_{v\in\mathbb{R}^n} [T(x,v)-t]^+ p(v)dv. \qquad (8)$$

It follows from standard properties of convex functions that $g_\beta(x,t)$ is convex in x and t if $T(x,v)$ is convex in x for fixed v [3]. We can also note that for fixed t, the function (8) is the *binning yield-loss* discussed in [6], *i.e.*, a linear penalty on delays that exceed a timing budget. However, although the mean-excess delay and the binning yield-loss acheive different goals, they can be implemented using similar algorithms.

The following theorem, due to Rockafellar and Uryasev [13] relates (7) to the mean-excess delay.

THEOREM 3.1. *If v is a continuous random vector, then the minimum value of $g_\beta(x,t)$ over t is equal to the mean-excess delay:*

$$m_\beta(x) = \inf_{t\in\mathbb{R}} g_\beta(x,t).$$

In effect, this theorem states that the mean-excess delay can be minimized by minimizing the function $g_\beta(x,t)$ over the design variables x and the additional variable t.

Another interesting fact is that the minimizer of (7) over t is the β-quantile of the distribution $T(x,v)$:

$$q_\beta(x) = \inf\{t \,|\, g_\beta(x,t) = m_\beta(x)\}$$

This gives us the resulting quantile for free as a by-product of the optimization.

In summary, for continuous distributions, we have the following properties:

1. $q_\beta(x) \le m_\beta(x)$
2. $m_\beta(x) = \inf_t g_\beta(x,t)$
3. $q_\beta(x) = \inf\{t \,|\, g_\beta(x,t) = m_\beta(x)\}$
4. $g_\beta(x,t)$ is jointly convex in t and x if $T(x,v)$ is convex in x.

These properties are illustrated in Fig. 1. Here, the minimum of $g_{0.95}(x,t)$ is $m_{0.95}(x)$ with a minimizer $q_{0.95}(x)$.

In the case of a discrete distribution, properties 1, 3 and 4 hold. However in this case, the mean-excess delay is defined to be $\inf_t\ g_\beta(x,t)$, and it can still be used as an upper bound on the quantile.

Using the properties above, we can now reformulate (6) as:

$$\begin{array}{ll} \text{minimize} & g_\beta(x,t) \\ \text{subject to} & A(x) \le A_{\max} \\ & P(x) \le P_{\max} \\ & x \ge 0 \end{array} \qquad (9)$$

with variables x and t.

Because the preceding discussion is not unique to circuit sizing, the mean-excess delay can be applied to other problems where the delay is a convex function of the design variables, for a fixed variation. Furthermore, the same type of analysis can be used to combine power and delay minimization.

4. MINIMIZATION ALGORITHM

In the case of a discrete probability distribution and model (2), problem (9) can be expressed as a geometric program. This can be solved using a general purpose solver such as Mosek [10], but for large problems and a large set of discrete variations, a specialized algorithm must be used. The case of continuous distributions can be approximated as a discrete case by taking a fixed number of samples. This is referred to as a *Sample Average Approximation* (SAA) of the problem.

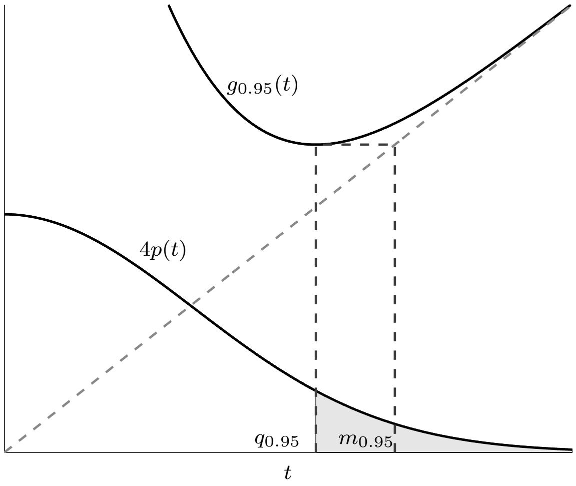

Figure 1: The pdf of the delay $T(x,v)$ for a Gaussian distribution is plotted above. $q_{0.95}(x)$ is the 95% quantile, and $m_{0.95}(x)$ is the 95% mean excess delay. The function $g_{0.95}(t)$ is plotted to illustrate the properties in Section IIIA. The minimizer of $g_{0.95}(t)$ is $q_{0.95}$, and the minimum value, which is reflected across the diagonal line, is $m_{0.95}$.

We solve the sample average approximation of the problem (9) using the Analytic-Centering Cutting Plane Method (ACCPM) [1]. In this method, we start with a polyhedral set S_0 that contains the optimum gate size. At each step of this method, the approximate "center" of the polyhedron is found, and the set of possible solutions S_k is reduced by adding a *cutting plane* through this center. The cutting plane *cuts away* the part of the region where the optimum cannot lie, reducing the set of possible solutions. In the ACCPM, this is achieved as follows.

1. Find the analytic center x_k^c of the current set of possible solutions S_k using Newton's method
2. Evaluate the gradient $\nabla g_\beta(x,t)$ of the Mean-Excess Delay at the point x_k^c
3. Use the cutting plane to reduce the set of possible solutions:
$$S_{k+1} = S_k \cap \{x \,|\, \nabla g_\beta(x,t)(x - x_k^c) \le 0\}.$$

The analytic center of a set of linear inequalities is defined as the minimizer of the log-barrier function:

$$x_c = \operatorname*{argmin}_x \left\{ -\sum_{i=1}^m \log(b_i - a_i^T x) \right\}$$

where the vectors a_i and scalars b_i define the cutting planes. This centering process is done in an effort to maximize the region that will be cut away by the cutting plane.

5. RESULTS

We sized the ISCAS '85 benchmark circuits using 3 different sizing methods:

Table 1: Sizing Results - Size independent variations

	c432		c499		c1355		c1908	
Method	$q_{0.95}$	d_{nom}	$q_{0.95}$	d_{nom}	$q_{0.95}$	d_{nom}	$q_{0.95}$	d_{nom}
Nominal	3.9ns	2.9ns	.73ns	.50ns	.79ns	.57ns	1.9ns	1.3ns
Padded	3.9ns	2.9ns	.73ns	.50ns	.79ns	.57ns	1.9ns	1.3ns
MED	3.9ns	2.9ns	.68ns	.51ns	.72ns	.58ns	1.8ns	1.3ns

	c2670		c3540		c5315		c7552	
Method	$q_{0.95}$	d_{nom}	$q_{0.95}$	d_{nom}	$q_{0.95}$	d_{nom}	$q_{0.95}$	d_{nom}
Nominal	.82ns	.63ns	1.5ns	1.0ns	1.1ns	.89ns	.96ns	.77ns
Padded	1.3ns	1.1ns	1.5ns	1.0ns	1.1ns	.89ns	.95ns	.77ns
MED	.79ns	.69ns	1.4ns	1.0ns	1.0ns	.93ns	.93ns	.83ns

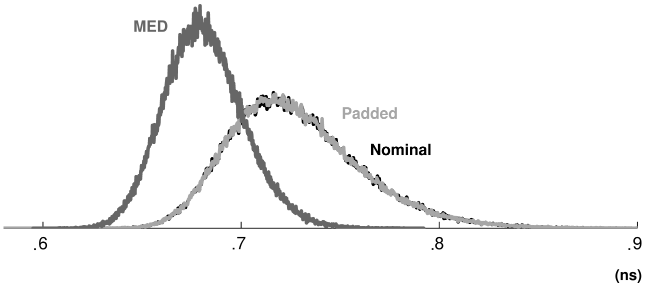

Figure 2: The delay pdfs of benchmark c1355 are plotted for each of the three sizing methods. The MED sizing gives the best pdf, followed by the padded delay sizing and the nominal sizing, which are nearly identical.

1. Nominal Sizing: variations are ignored
2. Padded Delay Sizing: a conservative robust sizing method [11]
3. Mean Excess Delay Sizing.

The standard deviations of each random variable were set to 0.25 and are split evenly between the "die-to-die" and "gate-by-gate" variations. These variations are made to be independent of each other, and they are assumed to be Gaussian. The nominal and padded delay sizing problems were solved using the general purpose commercial solver MOSEK [10], and the mean-excess delay sizing was done using the ACCPM with 2000 samples of the distribution.

With the size independent variations in (2), the Mean-Excess Delay Sizing always results in a better quantile than the nominal or padded delay sizings, but the magnitude of this difference depends on the structure of the circuit. This is summarized in Table 1. Here, $q_{0.95}$ denotes the 95% quantile and d_{nom} denotes the delay associated with the nominal case, where the variations are ignored. In the case of c432, the numbers were nearly identical for each of the three sizings, however in the case of the larger c1355, the difference was approximately 10%.

Table 2: Sizing Results - Size Dependent Variations

	c880		c1355		c2670		c3540	
Method	$q_{0.95}$	d_{nom}	$q_{0.95}$	d_{nom}	$q_{0.95}$	d_{nom}	$q_{0.95}$	d_{nom}
Nominal	.86ns	.49ns	.73ns	.56ns	.70ns	.64ns	1.2ns	1.1ns
Padded	.73ns	.50ns	.62ns	.58ns	.67ns	.66ns	1.1ns	1.1ns
MED	.73ns	.50ns	.61ns	.58ns	.66ns	.64ns	1.1ns	1.1ns

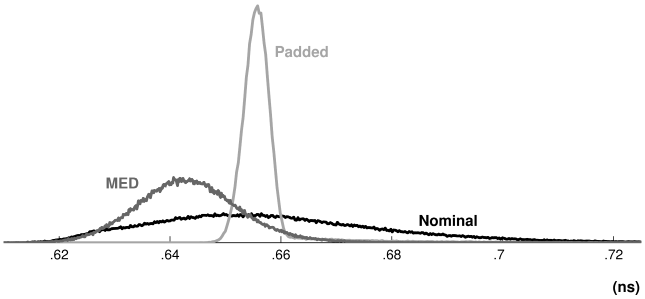

Figure 3: The delay pdfs of benchmark c2670 with size dependent variations are plotted for each of the three sizing methods. The MED gives the best delay distribution, followed by the padded delay sizings. These methods give very good results with significant gains over the nominal.

With the gate-size dependent variations in (3) the problem is no longer convex, and it loses the tractability that is associated with convex problems. However, a few runs were made to compare these methods under this model. We use the setting $\alpha = .3$ and the variations were set to be the same as above. In this case, the MED sizing can still give better results (see Table 2), however the improvements on the quantile are not as large, dropping to 2%. A closer look at the pdfs in Fig. 3 shows that although the 95% quantiles are competitve, the pdf of the MED sizing has a significantly faster average delay than the padded delay sizing.

There are three conclusions that can be drawn from these results. First, there is a significant gain when the statistical problem is solved over the nominal problem. With size independent variations, this can be up to 10% and with size dependent variations, the difference grows to 15%. Secondly, the padded delays usually give results that are similar to the nominal sizing, and in our experieince, we find this approximation to be useful when there is a large spread in path delays, such as large adder circuits. In contrast, the padded delay method does an excellent job when a size-dependent variation model is used, but there is still some improvement that can be made by using the MED sizing.

The runtime of this method scales similarly to the interior point solver used to perform the nominal and padded delay sizings when the number of samples is fixed. This is because both methods use Newton's method at each iteration to solve the subproblems. Furthermore, the same techniques that are used to improve scalability in the nominal sizing, as in [8], might apply to MED sizing with SAA sampling.

6. CONCLUSION

In this paper we introduced the mean excess delay as a statistical measure for circuit delay in the presence of random parameter variations. The β-mean excess delay is defined as the expected delay of the circuits whose delay exceed the β-quantile. Thus, minimizing mean excess delay indirectly reduces the β-quantile. This technique is known in finance as conditional value-at-risk optimization [13]. Compared with other popular methods (such as the padded delay sizing method [11]), MED sizing has the important advantage that it takes into account the type of correlation in the gate delay variations. The results show a significant improvement can be made by using the mean-excess delay over the padded and nominal methods. Future work will aim to improve the scalability of the algorithm and to study the problem with the model (3).

7. ACKNOWLEDGEMENTS

Financial supports from SRC contract 2006-TJ-1460 and NSF Awards CCF-0528583 and ECS-0524663 are gratefully acknowledged. We also thank the reviewers for their helpful suggestions.

8. REFERENCES

[1] D. S. Atkinson and P. M. Vaidya. A cutting plane algorithm for convex programming that uses analytic centers. *Mathematical Programming*, 69(1):1–44, July 1995.

[2] S. Boyd, S. Kim, D. Patil, and M. Horowitz. Digital circuit optimization via geometric programming. *Operations Research*, 53(6):899, 2005.

[3] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[4] H. Chang, V. Zolotov, S. Narayan, and C. Visweswariah. Parameterized block-based statistical timing analysis with non-gaussian parameters, nonlinear delay functions. In *Proc. Design Automation Conference*, pages 71–76, 2005.

[5] C. Chen, C. Chu, and D. Wong. Fast and exact simultaneous gate and wire sizing by lagrangian relaxation. *IEEE Trans. on Computer-Aided Design*, 18(7):1014–1025, 1999.

[6] A. Davoodi and A. Srivastava. Variability driven gate sizing for binning yield optimization. In *Proc. Design Automation Conference*, pages 959–964, 2006.

[7] L. He, A. Kahng, K. Tam, and J. Xiong Simultaneous buffer insertion and wire sizing considering systemic CMP variation and random L_{eff} variation. *IEEE Trans. on Computer-Aided Design*, 26(5):845–857, 2007.

[8] S. Joshi and S. Boyd. An efficient method for large-scale gate sizing. Submitted for publication, available at http://www.stanford.edu/~boyd/gatesizing.html.

[9] M. Mani and M. Orshansky. A new statistical optimization algorithm for gate sizing. In *Proc. IEEE Int. Conf. on Computer Design*, pages 272–277, 2004.

[10] MOSEK ApS. The MOSEK Optimization Tools Version 5.0. Available from http://www.mosek.com.

[11] D. Patil, S. Yun, S. Kim, A. Cheung, M. Horowitz, and S. Boyd. A new method for design of robust digital circuits. In *Proc. of ISQED*, pages 676–681, 2005.

[12] M. Pelgrom, A. Duinmaijer, and A. Welbers. Matching properties of MOS transistors. In *IEEE Journal of Solid-State Circuits*, 24(5):1433–1439, 1989.

[13] R. Rockafellar and S. Uryasev. Optimization of conditional value-at-risk. *Journal of Risk*, 2(3):21–41, 2000.

[14] J. Singh, V. Nookala, Z. Luo, and S. Sapatnekar. Robust gate sizing by geometric programming. In *Proc. Design Automation Conference*, pages 315–320, 2005.

[15] D. Sinha, N. Shenoy, and H. Zhou. Statistical gate sizing for timing yield optimization. In *Proc. Int. Conf. Computer-Aided Design*, pages 1037–1041, 2005.

[16] C. Visweswariah. Death, taxes and failing chips. In *Proc. Design Automation Conference*, pages 343–347, 2003.

[17] C. Visweswariah, K. Ravindran, K. Kalafala, S. Walker, and S. Narayan. First-order incremental block-based statistical timing analysis. In *Proc. Design Automation Conference*, pages 331–336, 2004.

[18] V. Wang, K. Agarwal, S. Nassif, K. Nowka, and D. Markovic. A design model for random process variability. To appear in ISQED '08.

[19] W. Wang, V. Kreinovich, and M. Orshansky. Statistical timing based on incomplete probabilistic descriptions of parameter uncertainty. In *Proc. Design Automation Conference*, pages 161–166, 2006.

[20] Y. Zhan, A. Strojwas, X. Li, L. Pileggi, D. Newmark, and M. Sharma. Correlation-aware statistical timing analysis with non-gaussian delay distributions. In *Proc. Design Automation Conference*, pages 77–82, 2005.