A General Probabilistic Framework for Worst Case Timing Analysis

Michael Orshansky and Kurt Keutzer
Department of Electrical Engineering and Computer Sciences
University of California, Berkeley

1 ABSTRACT

The traditional approach to worst-case static-timing analysis is becoming unacceptably conservative due to an ever-increasing number of circuit and process effects. We propose a fundamentally different framework that aims to significantly improve the accuracy of timing predictions through fully probabilistic analysis of gate and path delays. We describe a bottom-up approach for the construction of joint probability density function of path delays, and present novel analytical and algorithmic methods for finding the full distribution of the maximum of a random path delay space with arbitrary path correlations.

Categories and Subject Descriptors

J.6.1 [**Computer-Aided Engineering**]: Computer-aided design.

General Terms: Algorithms

2 INTRODUCTION

Over the years it has been widely acknowledged that the uncertainty about the true design and manufacturing conditions is a major cause of unnecessary over-design and resulting underperformance of circuits [1][2]. The sources of this uncertainty are manifold, and are due to the limitations of the actual design practices, uncertainty about the environmental design characteristics (cross-talk noise, temperature and supply voltage variation), and the inherent variation of the underlying process parameters. With the advance of deep sub-micron technologies, process variability and, in particular, intra-chip variation, has been increasing. This is due to various processing and device physics factors such as random dopant placement in the channel, spatially correlated and proximity-caused Lgate variation, and interconnect metal thickness variation [2].

The emergence of intra-chip parameter variability as a dominant source of uncertainty and circuit degradation requires a new set of approaches to circuit timing analysis, whose role is to guarantee that the predicted maximum clock speed is as close as possible to the actual (silicon) timing behavior. Industrial experience shows that the gap between the worst-case timing constraints predicted by the tools, and the final silicon performance is routinely greater than what can be tolerated and is sometimes as high as 30% [3].

What is wrong with the existing tools and approaches? Circuit-dependent parametric yield loss is predicted to become a key issue in nanometer silicon technologies [4]. The fundamental problem is that the standard timing techniques are incapable of accurately predicting parametric yield of a circuit due to their non-probabilistic formulation. One particular result of this failing is the well-known conservatism of traditional worst-case modeling techniques. We may distinguish at least two levels of conservatism. The first is the practice of defining the worst-case timing behavior of a cell by performing circuit analysis in SPICE that simultaneously sets all the device model parameters to their worst-case values. Several approaches to reduce this type of conservatism have been proposed [5]. When we move to the level of cell-based static timing analysis, an additional level of conservatism is created by the non-probabilistic delay computation of the traditional analysis. This conservatism is relatively new but rapidly growing in importance, and it arises due to the breakdown of key assumptions regarding the correlation between the timing responses of the various delay elements, implicit in the traditional worst-case timing tools.

In the past, several attempts have been made to introduce statistical computations into the domain of gate-level timing analysis. Hitchcock [6] describes a Monte-Carlo based technique for computing the distributions of gate delays. Jyu [7] proposes a faster approach that is capable of dealing with false paths. The deficiency of these techniques is that they are still computationally very expensive. A table look-up algorithm, considered by Berkelaar [8], is faster but fails to account for correlation between gate delays as well provide a way to compute the maximum of more than two variables.

In contrast to the earlier work, we propose an *analytical* theoretical framework for statistical timing analysis. Only the analytical (as opposed to computationally expensive sampling-based techniques, such as Monte-Carlo) methodology has a chance to be used in the timing analysis of real-sized circuits. The proposed approach is entirely probabilistic, seeking to construct the probability distribution of an achievable clock period for a given circuit. We show how to construct the joint probability density function (*jpdf*) of path delays, and, specifically, how to form the covariance matrix of *jpdf* starting from the statistical gate models, and statistical process variation models. The most difficult theoretical problem that has to be solved next is finding the distribution of the maximum of a multivariate vector of path delays with arbitrary covariance structure, described by the just-derived *jpdf*. We introduce a set of previously unavailable analytical estimates and algorithmic solutions that allow constructing tight upper and lower bounds on the distribution of the maximum.

We begin in the next section by motivating the need for a new probabilistic approach. Section 4 describes a modeling strategy for derivation of the joint probability density function of individual delay elements (both gates and wires) and path delays. Then, in section 5.1, the bounds for the distribution of the maximum of path delays, and, in section 5.2, an algorithm to compute them, are developed. Finally, in section 6, the circuit results and a set of quantitative comparisons are provided.

DAC 2002, June 10-14, 2002, New Orleans, Louisiana, USA.

3 WHY TRADITIONAL WORST-CASE TIMING IS OVERLY-PESSIMISTIC

Before we dive into the technical details of the new approach, let us look at why the traditional worst-case timing analysis is bound to result in an unreasonable level of conservatism. Let us for simplicity consider the gate-level static timing analysis. For the sake of simplicity we ignore the orthogonal issue of false paths. Worst-case static timing analysis proceeds by setting each gate to its worst-case timing value and performing a longest path computation to arrive at the worst-case critical path delay. The assumption that is implicit in this approach is that delay elements (gates and wires) are perfectly correlated with each other. The assumption of perfect correlation is reasonable only if the following two conditions hold: (a) intra-chip variation is negligible, and (b) the sensitivity of delay response of *each gate* and *each wire segment* to variation in *every parameter* is perfectly correlated with all the other delay elements (e.g. *both* gate and wire delays). We already know that with the rise of intra-chip variation the first assumption is getting less and less reliable. It is also known that the delay response of the different cells is not perfectly correlated [9]. Lastly, with the rise of the contribution of interconnect delay the fact of its poor correlation with gate delays cannot be ignored.

The failure to consider the validity of the assumptions makes the probability of finding a manufactured chip with the characteristics, assigned to it by the worst-case timing analysis, extremely small. The majority of manufactured chips exhibit a significantly higher performance, as confirmed by practice [3].

4 A PROBABILISTIC FOUNDATION FOR TIMING ANALYSIS

The fundamental problem of the traditional approach to worst-case timing analysis is that it is essentially formulated in a non-probabilistic manner. *The delays of gates, and later, of the paths are treated as fixed numbers, not random variables.* The inherently probabilistic problem is reduced to a purely arithmetical one, and once this transition is made, the ability to probabilistically quantify the likelihood of timing estimates is irreversibly lost. A different formulation of the timing problem is required that would do justice to the probabilistic nature of the problem. Such a completely general formulation is now advanced.

4.1 Problem Formulation

The clock cycle of a chip is constrained by the maximum path delay, $\max\{D_1...D_N\} \le T_{clock}$, where D_i is the delay of the i_{th} path in the circuit. The delay of each path is a random variable, described by a probability distribution. Because $\{D_1...D_N\}$ is a random vector, the value of $\max\{D_1...D_N\}$ is also a random variable. Then, in order to estimate the statistical properties of the chip's timing, we must find the distribution of $\max\{D_1...D_N\}$. The cumulative probability function of $\max\{D_1...D_N\}$ is given by $F(t) = P\{\max\{D_1...D_N\} \le t\}$, or equivalently:

$$F(t) = P\{D_1 \le t, D_2 \le t, ..., D_N \le t\} \qquad (1)$$

where *F(t)* is the cumulative probability function defined over the path delay probability space. In general, we can find the cumulative probability function by direct integration:

$$F(t) = \int_{-\infty}^{t} (N-1) \int_{-\infty}^{t} f(D_1, D_2, ...D_N) dD_1 dD_2 .. dD_N \qquad (2)$$

where $f(D_1, D_2, ...D_N)$ is the joint probability density function (*jpdf)* of $\{D_1...D_N\}$. Unfortunately, the direct evaluation of an *N*-dimensional integral for an arbitrary $f(D_1, D_2, ...D_N)$ is extremely difficult for large *N*.

Given that it is impractical to solve (2) directly for large *N*, we are faced with the task of finding the distribution of $\max\{D_1...D_N\}$ by some other means. While the exact analytical expressions for the distribution of $\max\{D_1...D_N\}$ are still not available, in section 5, we introduce a set of tight analytical bounds for the distribution of $\max\{D_1...D_N\}$. But, first, we describe the derivation of the *jpdf* of path delays, and specifically, show how to construct the path delay covariance matrix on the basis of individual delay elements (gates and wires).

4.2 Path Delay Distribution

We limit our analysis to a combinational circuit containing *N* paths. The analysis can be easily extended to cover sequential circuits. A path is defined by a set of gates and wires that this path traverses, but let us for simplicity talk about gates only, with an understanding that the wire delays can also be covered by the notation introduced below. Let there be a gate set *GS* of gates belonging to a path p_i: $GS\{p_i\} = \{g_i^1 ... g_i^{m_i}\}$, where m_i is the number of gates along the path *i*. In the presence of process variability and other sources of circuit variation, delay of each gate, and thus, each path, can be described by a random variable. Together, all path delays form a random vector $D = \{D_1...D_N\}$, whose joint probability density function, and especially, covariance structure, we want to establish.

We assume that, as is common practice, a Gaussian distribution best describes the process variation. Then, under the mild assumptions that we consider below, the individual gate delay and path delay distributions are also Gaussian. The *jpdf* of the *multivariate normal* vector of path delays is fully characterized by the vector of means, $E\{D\}$, and the covariance matrix, Σ_D: $D \sim N(E\{D\}, \Sigma_D)$. The mean vector of path delay *jpdf* is simply the *N*-dimensional vector of nominal delays of each path, e.g. $\mu = \{E\{D_1\},...,E\{D_N\}\}$. The mean delay of a path is given by $E\{D_j\} = \sum E\{d_g(i)\}$, for $i \in GS(p_j)$, where $d_g(i)$ is the delay of the i_{th} gate of path *j*. The *NxN* covariance matrix, Σ_D, is fully characterized by the pair-wise covariance terms between path delays, $\Sigma_{(i,j)} = \mathrm{cov}\{D_i, D_j\}$, where D_i is the random delay of path *i*.

Path covariances $\Sigma_{(i,j)}$ can be constructively computed on the basis of pair-wise gate delay covariances. In the analysis that follows, we assume that statistical interactions between gate delays can be neglected in computing the total path delay response to process variations. This assumption is usually justified. Then,

$$\mathrm{cov}\{D_i, D_j\} = \sum_{k_i=1}^{m_i} \sum_{k_j=1}^{m_j} \mathrm{cov}\{d_g(i,k_i), d_g(j,k_j)\} \qquad (3)$$

where m_i is the number of gates along path *i*, and $d_g(i,k_i)$ is the delay of gate k_i of path *i*. This equation can also be modified to account for the effect of gate delay dependence on the output slew

of the previous gate within the path, by propagating variance through a path using a chain rule.

We now derive gate delay covariances. First, we introduce a general notation that will give the ability to include a wide range of gate delay dependences on the design-level, and ultimately, layout-level parameters. Let the attributes specifying a gate be contained in a vector L. Among other characteristics, L can incorporate the geometric properties of each gate (vertical/horizontal, proximity, left/right), gate sizes, threshold voltages, and so on. Also, let M be subset of the elements of the vector L that are affected by the process and manufacturing variations, for example, $M = \{L_{gate}, V_{th}\}$. Then, M is a random vector, and for simplicity we assume that it has a zero mean vector, $M \sim N(0, \Sigma_M)$. In other words, L would describe the nominal values of gate parameters, while M would describe the deviations from the nominal.

Let the delay of a gate be given by an arbitrary function $d_g = f(L)$. In order to establish an expression for the pair-wise covariances of gate delays, we assume the linearity of delay response to the localized variation of process parameters. In other words, we assume that a first order Taylor expansion of the gate delay function is adequate. Then,

$$d_g = d_g(L_o) + \varphi(L_o)^T M \tag{4}$$

Here $\varphi(\cdot)$ is the Jacobian of the first-order derivatives of the delay function to M. For example, $\varphi(\cdot) = \left(\partial d_g / \partial L_g, \partial d_g / \partial V_{th}\right)$. Under the expansion of Eq. 4, the gate delay distribution is Gaussian. Letting $d = \dim\{M\}$, the covariance of two gate delays i and j is given by:

$$\operatorname{cov}\{d_g(i), d_g(j)\} = \sum_{t_i=1}^{d} \sum_{t_j=1}^{d} \varphi_{t_i} \varphi_{t_j} \operatorname{cov}\{M_{t_i}, M_{t_j}\} \tag{5}$$

Let us emphasize that our derivation has been completely general with respect to the particular nature of process variability and the types of delay elements we want to include in the analysis. Therefore, using this analysis, the *jpdf* (both the vector of means and the covariance matrix) of the random delay vector can reflect any design or layout dependence. And, again, while we talked only about 'gates', the wire delays as well as gate delays can be naturally accommodated into this analytical framework.

5 ANALYSIS OF GENERAL MULTINORMAL DISTRIBUTIONS

The probabilistic formulation of timing analysis advanced in section 4.1 requires us to estimate the distribution of $\max\{D_1 ... D_N\}$. The biggest obstacle to the analytical approach to statistical timing analysis, however, is the lack of a closed-form solution for $\max\{D_1 ... D_N\}$ of a system of random variables. While the various non-analytic ways of addressing this problem in the context of worst-case timing analysis have been explored, the analytic estimates have not been available for $N > 2$. In this section we derive tight analytical approximations for the distribution of $\max\{D_1 ... D_N\}$.

5.1 Finding Bounds of Gaussian Processes

It must be stated clearly that finding the exact distribution of the maximum of a multivariate distribution with an arbitrary covariance structure is an open mathematical problem. Only for certain types of covariance matrices such a solution can be found. Since the covariance of our multivariate path delay vector is unlikely to have any special symmetry, we can only derive bounds of the form $LB \le f(\max\{D_1 ... D_N\}) \le UB$, where LB and UB are correspondingly the lower and upper bounds. We establish the bounds by employing the powerful results from the study of the general Gaussian processes.

A brief explanation of notation used in this section is helpful. For a random variable X, EX is the expected (mean) value of X, and $EX^2 = E(X^2)$ is the variance of X. Note also that we use $\max\{X_1 ... X_N\}$ and $\max\{X\}$ interchangeably.

The first step in approximating the distribution $f(\max\{X\})$ is to find the location of the expected value of $\max\{X\}$ - $E\max\{X\}$. The following theorem is the fundamental result of the theory of Gaussian processes, and is key to our analysis [10].

Theorem 1. *Let X and Y have centered multivariate normal distributions, such that $EX_i^2 = EY_i^2$ for all i, j, and $E(Y_i - Y_j)^2 \le E(X_i - X_j)^2$. Then for all real λ*

$$P\{\max\{Y\} > \lambda\} \le P\{\max\{X\} > \lambda\}$$

In other words, for two multivariate distributions with constant variance, *the maximum of a more correlated distribution is stochastically smaller than the maximum of a less correlated distribution.* (Note that if $E(X_i - X_j)^2 \ge E(Y_i - Y_j)^2$, $\operatorname{cov}(X_i, X_j) \le \operatorname{cov}(Y_i, Y_j)$, i.e. Y has a more correlated distribution.) The importance of this inequality lies in the fact that it allows to deduce inequalities for multivariate normal distributions with complex covariance structures by comparing them with simpler distributions. First, a corollary of Theorem 1 for the relation between the expected maximums [10]:

Corollary 1. *Under conditions of Theorem 1*

$$E\max\{Y\} \le E\max\{X\}$$

$$\textit{Proof}: E\max\{Y\} = \int_0^\infty P\{\max\{Y\} > \lambda\} d\lambda - \int_{-\infty}^0 P\{\max\{Y\} < \lambda\} d\lambda$$

$$\le \int_0^\infty P\{\max\{X\} > \lambda\} d\lambda - \int_{-\infty}^0 P\{\max\{X\} < \lambda\} d\lambda = E\max\{X\}$$

Corollary 2. *If X and Y have centralized multivariate normal distributions, such that $EX_i^2 = EY_i^2$ for all i, j, and $\operatorname{cov}(X_i, X_j) = 0$ while $\operatorname{cov}(Y_i, Y_j) \ne 0$ for all i, j. Then*

$$E\max\{Y\} \le E\max\{X\}$$

The last corollary simply states that if two distributions have the same and constant variance, but one is a correlated distribution while the other is not, *the expected maximum of the un-correlated distribution establishes an upper bound for the expected maximum of the correlated distribution.* As we show in the next section, we can compute the exact value of the expected maximum of the un-correlated distribution. Combining it with Corollary 2, we get an upper bound on the expected maximum of the actual correlated distribution.

Theorem 1 can be extended to cover non-centralized multivariate normal distributions, e.g. distributions for which $EX_i \ne const$. Due to the lack of space, we present the result without a proof but it can be found in [11].

Theorem 2. *Let X and Y have multivariate normal distributions such that* $EX_i \neq const$, $EY_i \neq const$, *and* $EX_i = EY_i$. *Now, if* $EX_i^2 = EY_i^2$ $\text{cov}(X_i, X_j) = 0$, and $\text{cov}(Y_i, Y_j) \neq 0$ *for all i, j, then*

$$E\max\{Y\} \leq E\max\{X\}$$

Among other results, Theorem 1 leads to the useful lower bound on the expected maximum of a multivariate distribution. While parametric yield of a circuit, the percentage of chips that function properly, is determined by the upper bound, the lower bound helps to assess the amount of conservatism potentially contained in the upper bound [12]:

Theorem 3. *Let* $X = (X_1,...,X_N)$ *have a centered multivariate normal distribution, and let* $E(X_i - X_j)^2 \geq \delta^2$, *then*

$$E\max\{X\} \geq \frac{E|N(0,1)|}{\sqrt{8}}\delta\sqrt{\log_2 N},$$

where $E|N(0,1)| = 0.798$. For the sake of interpretation, note that $E(X_i - X_j)^2 = 2\sigma_X^2(1-\rho_{ij}) \geq \delta^2$, if the variance is constant. Thus, the lower bound is determined by the largest correlation between the elements of the multivariate normal vector.

So far our analysis has been limited to finding the expected value of the maximum. Ultimately, however, we would like to get an estimate of the probability distribution of $\max\{X_1...X_N\}$, or, at least, an estimate of some higher moments of its probability density function. This is due to the fact that in order to accurately predict parametric yield we need to be able to find the value of the maximum at different quantiles of the clock cycle distribution. Another fundamental inequality that comes from the study of general Gaussian processes helps us establish the bound for the distribution of the maximum. It states that the spread of $\max\{X_1...X_N\}$ about its mean is no worse than the spread in the distribution of X_i with largest variance. Moreover, the distribution of the maximum above its mean value is such that the likelihood of a deviation is bounded from above by the probability of the same deviation for a centralized normal variable with the variance equal to the largest variance [14].

Theorem 4. *Let* $X = (X_1,...,X_N)$ *have a centralized multivariate normal distribution. Let* $S = \max(X_i)$ *and* $\sigma_{\max}^2 = \max(\text{var}(X_i))$, *then for any* $r \geq 0$,

$$P\{(S - E\max\{X\}) > r\} \leq P\{N(0,\sigma_{\max}^2) \geq r\}$$

While Theorem 4 is formulated for a centralized distribution, the claim can be extended to an arbitrary mean vector. Due to the lack of space, we present the result without a proof but it can be found in [11].

Theorem 5. *Let X have a multivariate normal distribution such that* $EX_i \neq const$. *Then, under conditions of Theorem 4,*

$$P\{(S - E\max\{X\}) > r\} \leq P\{N(0,\sigma_{\max}^2) \geq r\}$$

Using this theorem we can directly establish the conservative bounds for the quantiles of the distribution of $\max\{X\}$.

Corollary 3. *Under conditions of Theorem 5, the value of* $\max\{X\}$ *at the* k_{th} *percentile of the distribution is:*

$$\max\{X\} \leq E\max\{X\} + z_k\sigma_{\max}$$

where z_k *is the value of the standard normal at the* k_{th} *percentile of the normal distribution.*

This result finally gives us all the necessary tools to bound $\max\{X\}$ of a general multivariate probability distribution. Later we provide some examples for the use of this result in estimating the parametric yield of a circuit.

5.2 An Algorithm for Exact Evaluation of Expected Maximum

In order to use the result of Corollary 3 to conservatively estimate the clock cycle, we first need to find $E\max\{D\}$ for a general path delay vector $D = \{D_1...D_N\}$. We know from Corollary 2 that the expected (mean) maximum of an un-correlated distribution establishes an upper bound for the expected (mean) maximum of the correlated distribution. In this section we describe an algorithm for exact evaluation of $E\max\{X\}$ for an un-correlated multivariate normal distribution with arbitrary mean and variance vectors.

The algorithm to find $E\max\{X\}$ is based on the following probabilistic idea: It is possible to find a point T_1, such that in the repeated drawings of N standard normal random variables, on average, at least one random variable out of N would exceed it. Because T_1 is the point, which is *guaranteed to be exceeded* by at least one variable, $T_1 \leq E\max\{X\}$. Let us now consider the single variable that is guaranteed to exceed T_1. We know that for this variable it is possible to find $T_1 \leq T_2$, such that $P\{\max\{X\} \geq T_2\} = P\{\max\{X\} \leq T_2\} = 1/2$. Then, by definition $T_2 = median(\max\{X\})$. Since by Theorem 4 the distribution of $\max\{X\}$ approaches normal distribution, the median approaches the mean, and we finally have $E(\max\{X\}) \to T_2$.

Let us consider a point T_2, and the probability p_i that a random variable X_i, distributed as $N(\mu_i, \sigma_i^2)$, would exceed it. Then,

$$p_i = p_i(T_2, \mu_i, \sigma_i) = P\{X_i > T_2\} = 0.5 - \Phi((T_2 - \mu_i)/\sigma_i) \quad (6)$$

where $\Phi(t) = 1/\sqrt{2\pi}\int_0^t \exp(-z^2/2)dt$ is the Laplace function.

For each component of the random vector $X = (X_1,...,X_N)$, we create a mapping to a Bernoulli random variable, whose probability of success, - the probability of X_i exceeding T_2, - is p_i. Let $n(N) = Z_1 + ... + Z_N$ denote the number of successes among N Bernoulli variables. The expected value of the number of successes is

$$E\{n(N)\} = EZ_1 + ... + EZ_N = \sum_{i=1}^{N} p_i \quad (7)$$

since for a Bernoulli variable $EZ_i = p_i$. As was observed above, in order to find $E\max\{X\}$, we need to find the point, at which the expected number of random variables exceeding it is ½. To find this point, T_2, we must solve the equation:

$$E\{n(N)\} = \sum_{i=1}^{N} p_i(T_2, \mu_i, \sigma_i) = 1/2 \quad (8)$$

with $p_i(T_2, \mu_i, \sigma_i)$ given by Eq. 6. If X is a centralized multivariate distribution with constant variance, we have $p_i = const$, and the equation simplifies to $N \cdot p(T_2, \mu, \sigma) = 1/2$. In this case, a closed form solution can be found with

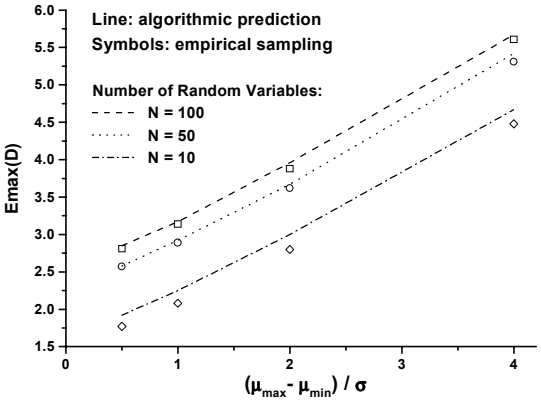

Figure 1. The algorithm is reasonably accurate for different ranges of the mean. Variance is fixed in this example. The results are shown for several values of $(\mu_{\max} - \mu_{\min})/\sigma$.

$E\max\{X\} = \eta$, where $\Phi(\eta) = (2N-1)/2N$. For X with a non-constant mean vector and/or non-constant variance, the equation (8) has to be solved iteratively.

The algorithm has been implemented in Matlab. We verified its accuracy by predicting $E\max\{X\}$ for different configurations of mean and variance vectors. The true value of $E\max\{X\}$ was found by generating sufficiently large random data sets. First we verified accuracy across the range of mean values by varying the ratio of $R = (\mu_{\max} - \mu_{\min})/\sigma$, under fixed variance, as shown in Figure 1. The algorithm always generates slightly conservative estimates but the accuracy improves, as N gets larger. For N=50 and N=100, the conservatism is very small 1-2%. The conservatism gets larger for smaller N and smaller R: for N=10 and R=0.5, it is about 8%. Because our approach is targeted for use with large N, this is an acceptable behavior of the algorithm.

We also considered accuracy of the algorithm for three different variance relations among the elements of the random vector, as shown in Figure 2: (a) homogenous (constant) variance,

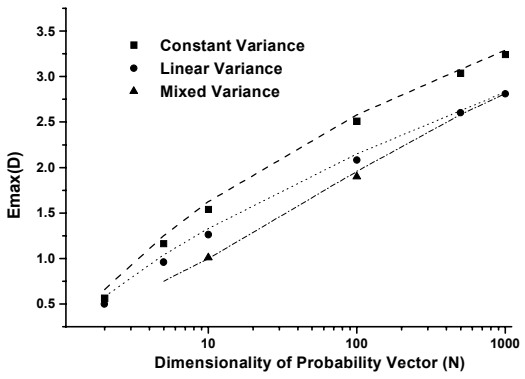

Figure 2. The algorithm is also accurate for distributions with different variance profiles. The mean is assumed constant.

(b) linear reduction of variance, and (c) a mixture of high and low variance. This time we assumed identical mean values. Accuracy is again very reasonable, as shown in Table 1. The values of $E\max\{X\}$ grow roughly logarithmically with N. $E\max\{X\}$ gets larger for the distributions with the larger number of elements with high variance (e.g. the constant variance distribution).

Table 1. The true values of $E\max\{X\}$, and the error of the algorithm's prediction. The algorithm is always conservative.

N	Constant Variance		Linear Variance	
	$E\max\{X\}$	Error (%)	$E\max\{X\}$	Error (%)
100	2.51	2.8	2.08	3.3
500	3.04	1.4	2.60	0.9
1000	3.24	1.5	2.81	0.7

6 RESULTS AND COMPARISONS

We now consider a simple circuit example that will illustrate several features of the probabilistic framework for worst-case timing analysis proposed in the previous sections. To highlight the important features of the approach, we constructed a topologically simple circuit that consists of a repeated pattern of 4 uniquely constructed paths. The four unique paths have similar but non-identical mean delays, $E\{D\}$. Because paths contain different number and types of cells, the variance of path delays, σ_D^2, is also different. (This way, path 4 whose $E\{D_4\} < E\{D_1\}$ may be stochastically slower than path 1, because $\sigma_4 > \sigma_1$.)

We use SPICE to analyze the circuit for a generic 0.18um CMOS technology. To evaluate the statistical properties of the circuit, we superimpose a variation of the gate length (L_{gate}) and the threshold voltage (V_{th}) on the nominal technology values. Both L_{gate} and V_{th} are assumed to be normally distributed; the standard deviations are given by $3\sigma_L = 0.15L_o$ and $3\sigma_{Vth} = 0.1V_{tho}$. We assume that both inter-chip and intra-chip components of variation are non-negligible, but that intra-chip component dominates. This may be expressed by setting $cor(L_i, L_j) = cor(V_{thi}, V_{thj}) = 0.3$. Equations 3-5 are used to estimate gate delay covariance, using SPICE-evaluated delay sensitivities. Path variance and covariance is then computed using Eq. 5.

Table 2. We consider a set of paths with different means and variances. The differences are due to path compositions.

Path	Path composition	# gates	$E\{D\}$ (ps)	σ_D (ps)	D^{wc} (ps)
1	AND2+NAND2+XOR2 +NOR2+NOR4	5	435	24.2	597
2	AND2+4NAND2 +NOR2+NOR4	7	430	18.5	580
3	INV+AND2+NAND2 +XOR2+ NOR4	5	424	23.6	579
4	2XOR+NOR4	3	425	27.9	590

Path delay distribution is clearly correlated. We estimate the upper bound for $E\max D$ using the exact algorithm of section 5.2. We also estimate the lower bound on $E\max D$ using the result of Theorem 3, and compare the newly derived bounds with the predictions of traditional timing analysis. The standard timing would find the *average* clock period given by 435ps, the largest $E\{D\}$ in Table 2. The results of probabilistic estimation of T_{clock} are in Table 3. For N=16, the average T_{clock} is predicted to be bounded by $443ps \le ET_{clock} \le 473ps$. In this case, standard

timing would underestimate the mean of the correlated path delay distribution by 1.5-8%. For N=400, the error would be 2-15%.

On the other hand, standard timing overestimates the worst-case behavior of a circuit. The standard timing analysis would conclude that the clock period is given by the largest worst-case delay (D^{wc}) among the paths of Table 2, giving $T_{clock}^{wc} = 597 ps$. The results of probabilistic timing analysis are derived for the 98th percentile of T_{clock} distribution. Using the results of Corollary 3:

$$T_{clock}(k) \le E\max\{D\} + z_k \sigma_{\max D}$$

where z_k is the value of the standard normal at the k_{th} percentile of the distribution. From Table 2 we have $\sigma_{\max D} = 27.9 ps$. Then, probabilistic timing analysis would conclude, for N=16, that $T_{clk}(0.98) \le 529 ps$. Thus, it overestimates the worst-case timing behavior by 6-12%, Table 3.

Table 3. The upper and lower bounds on $E\max\{D\}$ of correlated path delay distribution, the probabilistic T_{clock}, and the error comparison.

N	Bounds on $E\max\{D\}$, ps		$T_{clk}(0.98)$ (ps)	T_{clock}^{wc} (ps)	$\frac{T_{clock}^{wc} - T_{clk}(0.98)}{T_{clock}^{wc}}$ (%)
	Lower	Upper			
16	443	473	529	597	11.5
40	444	483	539	597	9.8
100	445	492	548	597	8.4
400	446	504	560	597	6.3

The difference between the standard and probabilistic timing is significantly larger for circuits with more balanced mean and variance vectors, and with more levels of logic in the paths. In the analysis of another circuit example, we found the difference in the worst-case timing behavior to be 7-21% [11].

The probabilistic formulation is vital, however, for reasons other than the difference in the timing estimates considered above. Among other things, it allows us to reveal the parametric yield curve. The true distribution of T_{clock} is tighter than given by the standard timing analysis, and is shifted towards the slower clock periods, (Figure 3). When the difference in T_{clock} distributions is mapped onto the parametric yield curve (which is just the cumulative distribution function of T_{clock}), one sees that a dramatic drop in parametric yield occurs at the high end of the clock frequency range. For example, the standard analysis predicts 50% yield for chips running at, or faster, than the average T_{clock}, while the probabilistic analysis shows that the yield is, at most, 8%.

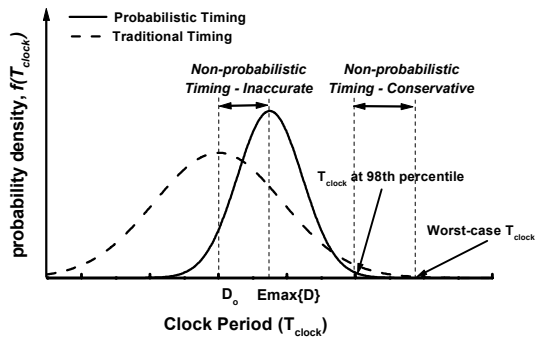

Figure 3. Traditional timing underestimates typical (mean) T_{clock} and overestimates worst-case T_{clock}.

Because the manufacturers of high-speed circuits get most of their revenue from the top percentile of the clock speeds, even the small decrease in parametric yield is very significant for them.

6 CONCLUSION AND FUTURE WORK

The innovation of this work is to define the problem of statistical timing analysis as the problem of finding the distribution of the maximum of circuit path delays with arbitrary covariance structure. The distribution of the maximum was estimated via the theoretical estimates and algorithms, which show a very good level of accuracy. Preliminary results indicate that traditional timing fails in several respects. It underestimates the value of the typical clock period, and overestimates the worst-case timing behavior, requiring expensive redesigns. It is also incapable of providing accurate parametric yield information. In the future, as the magnitude and complexity of process variation grows, we expect these deficiencies of non-probabilistic timing analysis to become even more prominent.

Overall, the results are encouraging, and indicate large potential for use of probabilistic timing analysis. More understanding of the computational behavior of the proposed algorithms for the analysis of large circuits is required. The other direction for substantial improvement is to further tighten the bounds on the probability distribution of the maximum of an arbitrary multivariate space. We believe that the development of such tighter bounds will show that the difference between the standard and probabilistic timing estimates is even greater.

7 REFERENCES

[1] Boning, D., and Nassif, S., "Models of Process Variations in Device and Interconnect", in *Design of High-Performance Microprocessor Circuits*, A. Chandrakasan (ed.), 2000.

[2] M. Orshansky, "Increasing Circuit Performance through Statistical Design Techniques," in *Closing the Gap Between ASIC and Custom*, Kluwer, D. Chinnery, K. Keutzer (ed.), 2002.

[3] Chinnery, D., and Keutzer, K., "Closing the Gap Between ASICs and Custom", *Proc. of DAC*, 2000.

[4] Semiconductor Industry Association, *International Technology Roadmap for Semiconductors*, 2001.

[5] S. Nassif, "Statistical worst-case analysis for integrated circuits," *Statistical Approaches to VLSI*, Elsevier Science, 1994.

[6] R. Hitchcock, "Timing Verification and the Timing Analysis Program", *Proc. of DAC*, 1982.

[7] H.-F. Jyu, S. Malik, S. Devadas, K. Keutzer, "Statistical timing analysis of combinational logic circuits," *IEEE Trans. on VLSI Systems*, vol.1, (no.2), June 1993. p.126-37.

[8] M. Berkelaar, "Statistical Delay Calculation", *International Workshop on Logic Synthesis,* 1997.

[9] S. Zanella et al, "Statistical Timing Macromodeling of Digital IP Libraries", *Workshop on Statistical Metrology*, 2000.

[10] R. Adler, *An Introduction to Continuity, Extrema, and Related Topics for General Gaussian Processes*, 1990, p. 49.

[11] M. Orshansky, *DAC 2002 Technical Materials*, www-device.eecs.berkeley.edu/~omisha/dac2002.htm

[12] D. Pollard, *A User's Guide to Measure Theoretic Probability*, Cambridge University Press, 2001.