Statistical Timing Analysis Based on a Timing Yield Model

Farid N. Najm
Department of ECE
University of Toronto
Toronto, Ontario, Canada

Noel Menezes
Strategic CAD Lab.
Intel Corporation
Hillsboro, Oregon, USA

ABSTRACT

Starting from a model of the within-die systematic variations using principal components analysis, a model is proposed for estimation of the parametric yield, and is then applied to estimation of the timing yield. Key features of these models are that they are easy to compute, they include a powerful model of within-die correlation, and they are "full-chip" models in the sense that they can be applied with ease to circuits with millions of components. As such, these models provide a way to do statistical timing analysis without the need for detailed statistical analysis of every path in the design.

Categories and Subject Descriptors

B.7 [**Integrated Circuits**]: Design Aids

General Terms

Algorithms, Design, Theory

Keywords

Statistical timing analysis, timing yield, principal components

1. INTRODUCTION

The yield of an integrated circuit (IC) is a complex function of a number of factors related to both design and manufacturing. Beyond issues of design centering [1, 2], which focuses mainly on tuning the manufacturing process, yield is also affected by circuit design. As part of circuit timing verification, one has to leave enough margin so that circuit delay variations do not affect yield too adversely. We will focus on this part of the overall yield problem, referred to as *timing yield* or *circuit-limited yield* [3, 4]. Traditionally, this has been taken care of by using the right *worst-case file* [5] as part of timing or performance verification (typically, during static timing analysis). The worst-case files specify the values of transistor parameters for various *process corners*, including the nominal and various extremes of device behavior. A circuit is deemed to have *passed* the timing test if it meets the performance constraints for all worst-case files belonging to that process. However, this approach is becoming less feasible today, especially for high-performance chips. For one thing, it can be too conservative and does not provide the user with any quantitative feedback on the robustness of the design [5]; it is a pass/fail approach. In addition, this traditional approach cannot handle *within-die* statistical variations [4] (mismatch between devices on the same chip) which have become important in deep sub-micron processes [3].

Statistical techniques offer a better alternative approach; statistical transistor modeling techniques [6, 5] have been used for quite some time. Recently, due to the increased importance of within-die variations, there has been an increased interest in tackling the timing yield problem by employing statistical techniques as part of the circuit timing analysis step [4, 3, 7, 8, 9]. The aim is to extend traditional static timing analysis so that it takes into account statistical delay variations.

Within-die variations are of two types: *systematic* and *statistical*. The systematic type are due to spatial location of a certain feature on the die and due to the context of that feature in terms of the layout patterns around and above it. Techniques have been proposed [10] for taking into account this type of variations. However, *statistical* within-die variations have not been adequately addressed. A key contribution of this paper is to take within-die statistical correlations into account.

To be sure, some prior work has been done on this. In a number of cases [4, 9, 11, 12], it has been assumed that within-die variations are totally uncorrelated (so that path yields are multiplied to give the chip yield), an assumption which is not true in practice. In order to avoid making this assumption, one needs to express the correlations between within-die parameter variations with a model *that can be easily built from process data.* This point is key, and is hard to do - there are no published models, for instance, for how exactly the variations are correlated across the die as a function, say, of the distance between components. A model of correlations in terms of distance is mentioned and used in [13], but no details or data are given; it is not clear what shape the model should take nor how one would build it from process data. In [8], even though statistical within-die variations are *not* taken into account, a suggestion is made at the end as to how one may include them and take care of correlation by enforcing correlation between features that are in the same region of the layout. This theme was further developed recently where use was made of principal components analysis [14] or a quad-tree partitioning [15] to express a region-wise spatial correlation among within-die variations. Here too, it is not clear how one would identify these regions and how the model would be built from process data. Finally, since these methods depend on placement information, these types of timing analysis become final sign-off tools and are unusable during circuit design.

We propose an approach to take within-die statistical correlations into account with a model that can be easily be built from process data, and which can be applied pre-placement.

2. PROPOSED APPROACH

In our approach, we capture the within-die correlations using principal components analysis (PCA). This is not, by itself, necessarily an improvement over prior art because, as with previous methods, the coefficients of the PCA would have to be evaluated somehow from process data, and would depend on position in the layout. However, with this model in hand, we then develop an approach to estimate a lower bound on the timing yield which which does *not* require knowledge of the individual PCA coefficients, but requires only the *order* of the PCA (the number of terms). While estimating detailed correlation functions is hard to do from process data, estimating only the order of the PCA would seem to be much easier, and we will suggest ways in which this may be done. Since layout information is not required, this ap-

DAC 2004, June 7–11, 2004, San Diego, California, USA.

proach becomes applicable to the pre-layout phase, during circuit design and optimization.

Previously proposed techniques for statistical static timing analysis change the static timing flow so that one is propagating distributions of delay, instead of simply delay. Even though there may be ways of doing this efficiently, this does represent a significant change of methodology! For one thing, statistical cell models [16] would need to be built if one is to use a cell-level or block-level flow. In contrast, our approach does not propagate distributions. Instead, the result of our approach is a selection of a "device file" setting with which to run traditional static timing analysis, which is somewhere within the extremes of device behavior. For example, while the "nominal" device file may call for a setting of $\Delta L = 0$ (for channel length variations) and the "worst-case" file may call for a setting of $\Delta L = +3\sigma_L$, our approach aims to predict the value "k" such that if the setting of $\Delta L = k\sigma_L$ was used for all devices, and if the circuit timing is verified using traditional static timing analysis, then the circuit would give the desired timing yield. As a result, our approach preserves existing static timing methodology and only assumes the existence of statistical transistor models, which have been standard for some time.

We will do this by working with a "generic critical path" concept, in the style of [13], and examining the statistical properties of large ensembles of such paths. Due to the typically huge number of critical paths on chip, the *law of large numbers* [17] will come into play, and we will show that the actual number of paths "drops out" of the yield equation. We will be left with a yield lower bound expression that depends only on the various components of the variances (these will be explained below) and on the "device file setting". For a minimum desired yield, we will work backwards to find the required settings of device parameters.

3. PARAMETER MODEL

For a given circuit element or layout feature i, let its coordinates on the die be (x_i, y_i) and let $X(i)$, be a zero-mean Gaussian *random variable* (RV) that denotes the variation of a certain parameter of this element from its nominal (mean) value. Thus, for example, $X(i)$ may represent channel length variations of transistor i. Correlation between values of $X(i)$ at different locations on the die may be expressed by means of an autocorrelation function, but this is not a practical approach. Instead, it is standard practice [18] to express the correlation by first breaking up the variations into *die-to-die* and *within-die* components, as follows:

$$X(i) = X_{dd} + X_{wd}(i) \tag{1}$$

The die-to-die component X_{dd} is an *independent* zero-mean Gaussian RV that takes the same value for all instances of this element on a given die, irrespective of location. The within-die component $X_{wd}(i)$ is a zero-mean Gaussian which can take different values for different instances of that element on the same die. This leads to the following relationship between the variances:

$$\sigma^2(i) = \sigma^2_{dd} + \sigma^2_{wd}(i) \tag{2}$$

Then, the within-die component is further broken down into two components, a *systematic* component and a "random" component:

$$X_{wd}(i) = X_{wds}(x_i, y_i) + X_{wdr}(i) \tag{3}$$

where, for each i, the random component $X_{wdr}(i)$ is an *independent* zero-mean Gaussian. Notice that the use of the term *systematic* here is somewhat different than the mention that was made of it in the introduction. In the introduction, we distinguished between systematic and statistical variations. In this case, the systematic component of the variations is statistical. Unfortunately, this same term is used in the literature to denote two different things. Throughout the rest of the paper, the term "systematic" will denote statistical variations such as $X_{wds}(x_i, y_i)$. The systematic component $X_{wds}(x_i, y_i)$ contains an explicit dependence on location because it is usually taken to represent the extent of correlation across the die, and correlation is usually dependent on relative location. A similar relationship follows for the variances:

$$\sigma^2_{wd}(i) = \sigma^2_{wds}(x_i, y_i) + \sigma^2_{wdr}(i) \tag{4}$$

One way to express the systematic component of the within-die variations is to use a *principal components analysis* (PCA) [19] and write:

$$X_{wds}(i) = \sum_{j=1}^{p} a_{ij} Z_j \tag{5}$$

where Z_j are independent *standard normal* RVs (Gaussians with zero mean and unity variance) and where the coefficients a_{ij} are such that:

$$\sigma^2_{wds}(x_i, y_i) = \sum_{j=1}^{p} a^2_{ij} \tag{6}$$

The RVs Z_j correspond to underlying independent unobservable factors. The value of p and the coefficients a_{ij} represent the extent of correlation across the die. For example, if $p = 1$, then the within-die spatial correlation coefficient is 1, there is perfect correlation; a single underlying RV Z_1 determines the value of the systematic component of X_{wd} all over the die. A $p > 1$ allows for less than perfect correlation.

We will adopt the PCA expansion (5) as our "correlation model" for the within-die component. At first glance, this model appears hard to use because it seems to depend on knowledge of the values of all the a_{ij} parameters. These coefficients depend on layout and, in any case, it is not clear how one would compute them from process data. A brute-force PCA expansion of the millions of instances of a parameter on a chip would be impractical. However, it will be shown in the following sections that one can estimate a lower bound on the yield without having to know the values of the a_{ij} parameters. Instead, it will be sufficient to know: 1) the "order" (p) of the PCA expansion, and 2) the ranges (max and min) of the variance terms given above. In the paper, the crucial step in the analysis is in the transition from (27) to (28), where an expression for yield that depends on all the a_{ij} terms is transformed (using Cauchy's inequality) to one that depends *only* on the sum of all the a^2_{ij} terms, which is easily available as the variance (6).

An important question, however, is how p is to be estimated. We can name three ways in which this may be done. First, based on knowledge of the process, it may be possible to simply *identify* a number of underlying independent factors that are responsible for the systematic variations, such as specific equipment or process steps. Second, one may associate each Z_j with a certain spatial location on the die, such as was done recently in [14]. Thus, if the chip area is partitioned into, say, four quadrants, and if one has some sense about distances over which the autocorrelation functions die down, one may be able to make an estimation of p. Third, and this may be the most practical approach, we can measure yield for a certain parameter, from process data, and then use the formulas to be derived below for parametric yield to work backwards to compute a value of p.

4. PARAMETRIC YIELD MODEL

With the random parameter model given above, we now define the *parametric yield* for parameter X as:

$$Y(x) = \mathcal{P}\{X(i) \le x,\ i = 1, 2, \ldots, n\} \tag{7}$$

where n is the number of instances of this parameter on chip. Here, $X(i)$ is a *generic* parameter that may represent transistor channel length variations, threshold voltage variations, etc. In fact, $X(i)$ is *any* statistical quantity on chip that may be characterized by the parameter model introduced in section 3. When $X(i)$ is a simple parameter, such as channel length, then parametric yield is the probability that *all* device lengths on the die are less than some threshold x. We will later on below show how path delay can itself be viewed as a *parameter* with its own triplet of variances (σ^2_{dd}, σ^2_{wds}, σ^2_{wdr}) that we will relate to the underlying transistor parameter variances. This will allows us to express timing yield based on a parametric yield model. Thus, the material in this section, although focused on parametric yield, will actually be directly useful for computing timing yield. Note also that, although we are expressing yield as a function of an upper-bound constraint on the value of the parameter, our work can be extended to cover lower bound and/or interval constraints.

Since X_{dd} is an independent zero-mean Gaussian with variance σ^2_{dd}, then $Z_0 = X_{dd}/\sigma_{dd}$ is an independent standard normal RV (mean 0, variance 1), and the expression for $Y(x)$ can be expanded as:

$$Y(x) = \mathcal{P}\{\sigma_{dd} Z_0 + X_{wds}(x_i, y_i) + X_{wdr}(i) \le x,\ \forall i\} \tag{8}$$

We now recall a result from basic probability theory that will be used repeatedly in the paper. Let $\mathcal{A}$ be an arbitrary event, and X be an RV with a probability density function (pdf) $f(x)$. Then (see [17], pg. 85) we have:

$$\mathcal{P}\{\mathcal{A}\} = \int_{-\infty}^{+\infty} \mathcal{P}\{\mathcal{A} \mid X = x\} f(x) dx \tag{9}$$

This result is simply an extension to the continuous case of the simple fact that $\mathcal{P}\{\mathcal{A}\} = \mathcal{P}\{\mathcal{A} \mid \mathcal{B}\} \cdot \mathcal{P}\{\mathcal{B}\} + \mathcal{P}\{\mathcal{A} \mid \overline{\mathcal{B}}\} \cdot \mathcal{P}\{\overline{\mathcal{B}}\}$, where $\mathcal{B}$ is another event. Applying (9) to (8), and denoting by $\phi(\cdot)$ the pdf of the standard normal distribution, gives:

$$Y(x) = \int\limits_{-\infty}^{+\infty} \mathcal{P}\{X_{wds}(x_i, y_i) + X_{wdr}(i) \le x - \sigma_{dd} z,\ \forall i\} \phi(z) dz \tag{10}$$

Let $X_{wd} = \max\limits_{\forall i} \left(X_{wds}(x_i, y_i) + X_{wdr}(i)\right)$ and denote its cumulative distribution function (cdf) by $Y_{wd}(a) = \mathcal{P}\left\{X_{wd} \le a\right\}$, then:

$$\begin{aligned} Y(x) &= \int_{-\infty}^{+\infty} \mathcal{P}\left\{X_{wd} \le x - \sigma_{dd} z\right\} \phi(z) d(z) \\ &= \int_{-\infty}^{+\infty} Y_{wd}(x - \sigma_{dd} z) \phi(z) dz \qquad (11) \end{aligned}$$

which means that:

$$Y(x) = E\Big[Y_{wd}\left(x - \sigma_{dd} Z_0\right)\Big] \tag{12}$$

where $E[\cdot]$ is the mean or expected value operator. The bulk of the effort will now be directed at computing the cdf $Y_{wd}(a)$. We first consider the special case $p = 1$ separately, before covering the general case.

4.1 Special Case $p = 1$

In this case, $X_{wds}(x_i, y_i) = \sigma_{wds}(x_i, y_i) Z_1$, where Z_1 is an independent standard normal. Therefore:

$$\begin{aligned} &Y_{wd}(a) = \mathcal{P}\left\{\sigma_{wds}(x_i, y_i) Z_1 + X_{wdr}(i) \le a,\ \forall i\right\} \\ &= \int_{-\infty}^{+\infty} \prod_{i=1}^{n} \mathcal{P}\left\{X_{wdr}(i) \le a - \sigma_{wds}(x_i, y_i) z\right\} \phi(z) dz \qquad (13) \end{aligned}$$

where we have again made use of (9) and of the fact that $X_{wdr}(i)$ are independent. Since $X_{wdr}(i)$ is a zero-mean Gaussian with variance $\sigma^2_{wdr}(i)$, then:

$$\begin{aligned} Y_{wd}(a) &= \int_{-\infty}^{+\infty} \prod_{i=1}^{n} \Phi\left(\frac{a - \sigma_{wds}(x_i, y_i) z}{\sigma_{wdr}(i)}\right) \phi(z) dz \qquad (14) \\ &= E\left[\prod_{i=1}^{n} \Phi\left(\frac{a - \sigma_{wds}(x_i, y_i) Z_1}{\sigma_{wdr}(i)}\right)\right] \qquad (15) \end{aligned}$$

where $\Phi(\cdot)$ is the cdf of the standard normal. Let $\sigma_{wds0} = \min\limits_{\forall i}(\sigma_{wds}(x_i, y_i))$ and $\sigma_{wds1} = \max\limits_{\forall i}(\sigma_{wds}(x_i, y_i))$. Since $\Phi(\cdot)$ is monotonically increasing, then (14) leads to the lower bound:

$$\begin{aligned} Y_{wd}(a) \ge &\int_{-\infty}^{0} \prod_{i=1}^{n} \Phi\left(\frac{a - \sigma_{wds0} z}{\sigma_{wdr}(i)}\right) \phi(z) dz \qquad (16) \\ + &\int_{0}^{+\infty} \prod_{i=1}^{n} \Phi\left(\frac{a - \sigma_{wds1} z}{\sigma_{wdr}(i)}\right) \phi(z) dz \qquad (17) \end{aligned}$$

Let $\sigma_{wdr0} = \min\limits_{\forall i}(\sigma_{wdr}(i))$ and $\sigma_{wdr1} = \max\limits_{\forall i}(\sigma_{wdr}(i))$, and consider the possible ranges of values of a. If $a \ge 0$, then (16) is minimized at σ_{wdr1} and (17) is broken into two integrals, one of which is minimized at σ_{wdr1} and the other at σ_{wdr0}, so that:

$$\begin{aligned} Y_{wd}(a) \ge &\int_{-\infty}^{0} \Phi^n\left(\frac{a - \sigma_{wds0} z}{\sigma_{wdr1}}\right) \phi(z) dz \\ + &\int_{0}^{a/\sigma_{wds1}} \Phi^n\left(\frac{a - \sigma_{wds1} z}{\sigma_{wdr1}}\right) \phi(z) dz \\ + &\int_{a/\sigma_{wds1}}^{+\infty} \Phi^n\left(\frac{a - \sigma_{wds1} z}{\sigma_{wdr0}}\right) \phi(z) dz \qquad (18) \end{aligned}$$

and if $a \le 0$ then a similar analysis leads to:

$$\begin{aligned} Y_{wd}(a) \ge &\int_{-\infty}^{a/\sigma_{wds0}} \Phi^n\left(\frac{a - \sigma_{wds0} z}{\sigma_{wdr1}}\right) \phi(z) dz \\ + &\int_{a/\sigma_{wds0}}^{0} \Phi^n\left(\frac{a - \sigma_{wds0} z}{\sigma_{wdr0}}\right) \phi(z) dz \\ + &\int_{0}^{+\infty} \Phi^n\left(\frac{a - \sigma_{wds1} z}{\sigma_{wdr0}}\right) \phi(z) dz \qquad (19) \end{aligned}$$

If these lower bounds on $Y_{wd}(a)$ are used in (12) then the result would be a lower bound $Y_0(x)$ on the yield: $Y(x) \ge Y_0(x)$. These bounds are expected to be tight if the differences $(\sigma_{wds1} - \sigma_{wds0})$ and $(\sigma_{wdr1} - \sigma_{wdr0})$ are small. With a simple change of variables, as will be illustrated below, $Y_0(x)$ can be computed by numerical integration. If a yield of, say, better than 90% is desired, then one can set $Y_0(x) = 0.9$ and work backwards to get the value of the threshold x.

4.1.1 Illustration

For illustration purposes, it is instructive to consider the simplified special case when $\sigma_{dd} = \sigma_{wds}(x_i, y_i) = \sigma_{wdr}(i) = \sigma$, $\forall i$. In this case, it becomes possible to get exact solutions to (12), instead of lower bounds, as follows. Starting from (15), we get:

$$Y_{wd}(a) = E\left[\Phi^n\left(\frac{a}{\sigma} - Z_1\right)\right] \tag{20}$$

and, combining this with (12) (to see why it is valid to combine the two in this way, see pp. 164–165 of [17]) leads to:

$$\boxed{Y(x) = E\Big[\Phi^n\big((x/\sigma) - Z_0 - Z_1\big)\Big]} \tag{21}$$

In order to compute this, we use the definition of the expected value operator (as an integral) and with a change of variables of $u = \Phi(z_0)$ and $v = \Phi(z_1)$, we arrive at:

$$Y(x) = \int_0^1 \int_0^1 \Phi^n\left(\frac{x}{\sigma} - \Phi^{-1}(u) - \Phi^{-1}(v)\right) du dv \tag{22}$$

Plots of this parametric yield for different values of n are shown in Fig. 1. Notice that yield decreases for larger n, as expected.

4.2 The General Case $p \ge 1$

Here $X_{wds}(x_i, y_i) = \sum\limits_{j=1}^{p} a_{ij} Z_j$, where $\sum\limits_{j=1}^{p} a^2_{ij} = \sigma^2_{wds}(x_i, y_i)$, so that:

$$\begin{aligned} Y_{wd}(a) &= \mathcal{P}\left\{\sum_{j=1}^{p} a_{ij} Z_j + X_{wdr}(i) \le a,\ \forall i\right\} \qquad (23) \\ &= \int\limits_{z_1=-\infty}^{+\infty} \cdots \int\limits_{z_p=-\infty}^{+\infty} \prod_{i=1}^{n} P_i(a) \phi(z_1) \cdots \phi(z_p) dz_1 \cdots dz_p \qquad (24) \end{aligned}$$

where we have made use of (9) a number of times (p) and of the fact that $X_{wdr}(i)$ are independent, and where:

$$P_i(a) = \mathcal{P}\left\{X_{wdr}(i) \le a - \sum_{j=1}^{p} a_{ij} z_j\right\} \tag{25}$$

We know from basic probability theory that, if X is an RV and $a_1 \le a_2$ are two real numbers, then $\mathcal{P}\{X \ge a_1\} \ge \mathcal{P}\{X \ge a_2\}$. In the problem at hand, we have:

$$\sum_{j=1}^{p} a_{ij} z_j \le \left|\sum_{j=1}^{p} a_{ij} z_j\right| \le \sqrt{\sum_{j=1}^{p} a^2_{ij}} \sqrt{\sum_{j=1}^{p} z^2_j} \tag{26}$$

where the 2nd inequality follows from Cauchy's inequality [17]. Therefore:

$$\begin{aligned} P_i(a) &= \mathcal{P}\left\{(a - X_{wdr}(i)) \ge \sum_{j=1}^{p} a_{ij} z_j\right\} \qquad (27) \\ &\ge \mathcal{P}\left\{(a - X_{wdr}(i)) \ge \sqrt{\sum_{j=1}^{p} a^2_{ij}} \sqrt{\sum_{j=1}^{p} z^2_j}\right\} \qquad (28) \end{aligned}$$

which, using $\sum\limits_{j=1}^{p} a^2_{ij} = \sigma^2_{wds}(x_i, y_i)$, leads to:

$$\begin{aligned} P_i(a) &\ge \mathcal{P}\left\{X_{wdr}(i) \le a - \sigma_{wds}(x_i, y_i) \sqrt{\sum_{j=1}^{p} z^2_j}\right\} \\ &= \Phi\left(\frac{a - \sigma_{wds}(x_i, y_i) \sqrt{\sum_{j=1}^{p} z^2_j}}{\sigma_{wdr}(i)}\right) \qquad (29) \end{aligned}$$

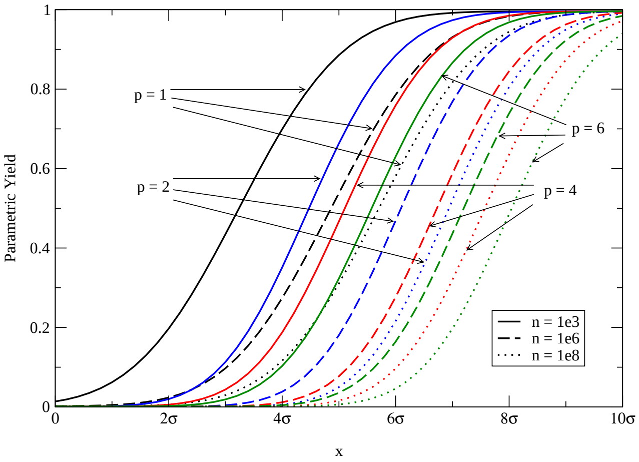

Figure 1: Parametric yield.

The transition from (27) to (28) is a key step and is fundamental to our contribution. A yield expression based on (27) would be very hard to use in practice because it requires knowledge of the individual a_{ij} coefficients. As previously mentioned, it is not clear how one would obtain these coefficients, which anyway are functions of layout, from the process data. However, estimation of the lower bound based on (28) would be quite feasible because it depends only on knowledge of the variance. It is precisely this step in the analysis that allows us to make yield estimation based on a generic critical path and without requiring layout information.

Plugging (29) into (24) gives:

$$\begin{aligned} Y_{wd}(a) &\geq E\left[\prod_{i=1}^{n} \Phi\left(\frac{a - \sigma_{wds}(x_i, y_i)\sqrt{\sum_{j=1}^{p} Z_j^2}}{\sigma_{wdr}(i)}\right)\right] \\ &\geq E\left[\prod_{i=1}^{n} \Phi\left(\frac{a - \sigma_{wds1}\sqrt{\sum_{j=1}^{p} Z_j^2}}{\sigma_{wdr}(i)}\right)\right] \end{aligned} \qquad (30)$$

where the 2nd inequality is true because $\sqrt{\sum Z_j^2} \geq 0$, and where σ_{wds1}, as before, is the largest $\sigma_{wds}(x_i, y_i)$. Let $Q_p \geq 0$ be an independent positive RV such that $Q_p^2 = \sum_{j=1}^{p} Z_j^2$, then Q_p^2 has the *chi-square* (χ^2) distribution with p degrees of freedom [17]. Therefore, we can replace the above right-hand-side by:

$$Y_{wd}(a) \geq E\left[\prod_{i=1}^{n} \Phi\left(\frac{a - \sigma_{wds1} Q_p}{\sigma_{wdr}(i)}\right)\right] \qquad (31)$$

$$= \int_0^{\infty} \prod_{i=1}^{n} \Phi\left(\frac{a - \sigma_{wds1}\sqrt{q}}{\sigma_{wdr}(i)}\right) f_{\chi_p^2}(q) dq \qquad (32)$$

where $f_{\chi_p^2}(\cdot)$ is the pdf of the χ^2 distribution with p degrees of freedom. As was done in the $p = 1$ case, a case-analysis based on the sign of a gives our final result. If $a \geq 0$ then:

$$\begin{aligned} Y_{wd}(a) \geq & \int_0^{a^2/\sigma_{wds1}^2} \Phi^n\left(\frac{a - \sigma_{wds1}\sqrt{q}}{\sigma_{wdr1}}\right) f_{\chi_p^2}(q) dq \\ & + \int_{a^2/\sigma_{wds1}^2}^{\infty} \Phi^n\left(\frac{a - \sigma_{wds1}\sqrt{q}}{\sigma_{wdr0}}\right) f_{\chi_p^2}(q) dq \end{aligned} \qquad (33)$$

where, as before, σ_{wdr1} and σ_{wdr0} are the largest and smallest $\sigma_{wdr}(i)$, respectively. If $a \leq 0$, then similarly:

$$Y_{wd}(a) \geq \int_0^{\infty} \Phi^n\left(\frac{a - \sigma_{wds1}\sqrt{q}}{\sigma_{wdr0}}\right) f_{\chi_p^2}(q) dq \qquad (34)$$

4.2.1 Illustration

Consider again the special case where $\sigma_{dd} = \sigma_{wds}(x_i, y_i) = \sigma_{wdr}(i) = \sigma$, $\forall i$. Then (31) leads to:

$$Y_{wd}(a) \geq E\left[\Phi^n\left(\frac{a}{\sigma} - Q_p\right)\right] \qquad (35)$$

and, combining this with (12), leads to:

$$\boxed{Y(x) \geq Y_0(x) = E\left[\Phi^n\left((x/\sigma) - Z_0 - Q_p\right)\right]} \qquad (36)$$

Plots of this parametric yield are given in Fig. 1, for various values of p and n. Notice that a larger p has the same effect as a larger n; more things can go wrong and the yield is lower.

4.3 Bounded Variations

Notice that, in the above expressions for yield, the yield decreases for larger n. One would somewhat expect this, but it is surprising to note that the yield approaches zero as n goes to infinity, for any combination of values of the three variances. This may also be seen in the above plots in Fig. 1. Even if the threshold is set at 10σ, there is still some parametric yield loss. This is somewhat non-physical, and arises due to the fact that we have assumed that the distribution of $X_{wdr}(i)$ is normal; recall that the normal distribution extends to $\pm\infty$ in both directions. In reality, one would expect process variations to be bounded by some upper and lower bounds. If a device somewhere deviates by large amounts, like 9σ or 10σ, then chances are there is a serious problem with that die, and that it would be lost due to other reasons, other than timing yield that is. Therefore, it is a good idea to limit the spread the cdf of $X_{wdr}(i)$ to some multiple of σ in order to avoid these non-physical effects at large n. In this section, therefore, we will use a *truncated normal* distribution for $X_{wdr}(i)$. For clarity of presentation, we will restrict the analysis to the illustrative special case introduced above wherein $\sigma_{dd} = \sigma_{wds}(x_i, y_i) = \sigma_{wdr}(i)$, $\forall i$. The analysis can be extended to the general case. Suppose, therefore, that $X_{wdr}(i)$ is bounded by $\pm k\sigma$, and let $\Phi_t(x)$ represent the cdf of the truncated standard normal, which is 0 for $x \leq -k$ and 1 for $x > k$.

We can plug $\Phi_t(\cdot)$ instead of $\Phi(\cdot)$ (for $X_{wdr}(i)$) into the above equations and plot the resulting yield integrals, as shown in Fig. 2. In this case the yield loss at higher n values is limited so that the 1e6 and 1e8 plots in each group are indistinguishable. This is to be expected, because the "tail" of the distribution has been cut off, and it is primarily the tail that causes the yield loss at very large n.

When working with a truncated normal, it is noteworthy that we can derive lower bounds on the yield that are independent of n. The derivations are not shown, for brevity, but lead to the following results. For the special case when $p = 1$, we have:

$$Y(x) \geq Y_0(x) = E\Big[\Phi\left((x/\sigma) - k - Z_0\right)\Big] \qquad (37)$$

A plot of this yield lower bound is shown in Fig. 2, in the $p = 1$ group. The lower bound is very tight, and is indistinguishable on the plot from the 1e6 and 1e8 curves in that group. In the general case of $p \geq 1$, we can show:

$$Y_0(x) = E\Big[U((x/\sigma) - k - Z_0)\, \chi_p^2\left(((x/\sigma) - k - Z_0)^2\right)\Big] \qquad (38)$$

where $U(x)$ is 1 for $x \geq 0$ and 0 for $x < 0$. Some plots of this yield lower bound are shown in Fig. 2. As before, the lower bound is very good, and is indistinguishable on the plot from the 1e6 and 1e8 curves in each group.

5. TIMING YIELD MODEL

In deep sub-micron CMOS, process-induced delay variations are due mainly to variations in the MOSFET threshold voltage (V_t) and effective channel length (L_e). Due to short-channel effects, V_t may depend on L_e (the so-called V_t roll-off effect), so that V_t and L_e are not independent variables. We capture this by assuming that V_t can be expressed as the sum of a term that depends on L_e and another independent term, so that, as RVs, we can express V_t and L_e of transistor i as follows:

$$\begin{aligned} L_e(i) &= E\left[L_e(i)\right] + L(i) \\ V_t(i) &= E\left[V_t(i)\right] + f(L(i)) + V(i) \end{aligned} \qquad (39)$$

where $E[\cdot]$ is the mean (expected value) operator. We are interested in the RVs $L(i)$ and $V(i)$, which are assumed to be *independent* of each other. We assume these RVs to be zero-mean Gaussians and break them up in the usual manner as:

$$\begin{aligned} L(i) &= L_{dd} + L_{wds}(x_i, y_i) + L_{wdr}(i) \\ V(i) &= V_{dd} + V_{wds}(x_i, y_i) + V_{wdr}(i) \end{aligned} \qquad (40)$$

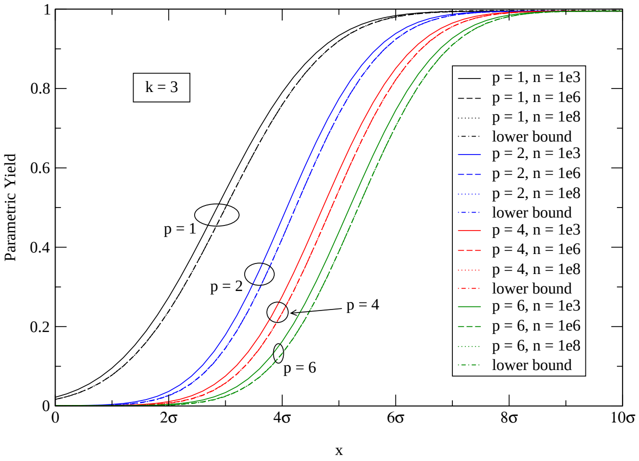

Figure 2: Parametric yield for $k = 3$.

(Of course, here V_{dd} is the die-to-die component of $V(i)$, and not the supply voltage.) We assume that the variances of $L(i)$ and $V(i)$, as well as the variance of their components, are known from the technology files:

$$\begin{aligned} \sigma_L^2(i) &= \sigma_{dd,L}^2 + \sigma_{wds,L}^2(x_i, y_i) + \sigma_{wdr,L}^2(i) \\ \sigma_V^2(i) &= \sigma_{dd,V}^2 + \sigma_{wds,V}^2(x_i, y_i) + \sigma_{wdr,V}^2(i) \end{aligned} \tag{41}$$

5.1 Gate Delay

We assume that, for a given chip design in a given technology, one can define a "nominal" representative logic gate, with appropriate output loading and input slope. For reasons that will become clear below, this gate should be typical of gates on critical paths in this technology. Due to the nonlinearity of the relationship between gate delay and transistor parameters, the mean value of gate delay does not necessarily coincide with its *nominal* value (the value corresponding to the case when $L(i) = 0$ and $V(i) = 0$). Furthermore, the distribution of gate delay would not necessarily be Gaussian. Simple experiments with HSPICE, however, reveal that this nonlinearity is not strong, at least not in 0.13μm CMOS. Therefore, we will ignore these complications for now, and simply assume that gate delay is a Gaussian with mean equal to its nominal value.

For all the transistors within a logic gate, we assume that their channel length variations are captured with a *single* RV $L(i)$ and their threshold voltage variations are captured with a *single* RV $V(i)$. Having ignored the nonlinearity between gate delay variations and the V_t and L_e variations, then if $D(i)$ is the deviation of the delay of logic gate i from it's mean (nominal) delay, we have:

$$D(i) = \alpha L(i) + \beta V(i) \tag{42}$$

where α and β are sensitivity parameters, with suitable units, that one can easily obtain from circuit simulation of a representative logic gate. Notice that, in general, $\alpha > 0$ and $\beta > 0$. (For a specific industrial 0.13μm process, we have found that for a minimum-sized inverter, $\alpha \approx 0.857$ps/nm and $\beta \approx 17.3$ps/V.) As a result, we can express the statistical variations in delay of gate i as:

$$D(i) = D_{dd} + D_{wds}(x_i, y_i) + D_{wdr}(i) \tag{43}$$

so that $D_{dd} = \alpha L_{dd} + \beta V_{dd}$, $D_{wds}(x_i, y_i) = \alpha L_{wds}(x_i, y_i) + \beta V_{wds}(x_i, y_i)$, and $D_{wdr}(i) = \alpha L_{wdr}(i) + \beta V_{wdr}(i)$. This leads to $\sigma_{dd,D}^2 = \alpha^2 \sigma_{dd,L}^2 + \beta^2 \sigma_{dd,V}^2$, $\sigma_{wds,D}^2(x_i, y_i) = \alpha^2 \sigma_{wds,L}^2(x_i, y_i) + \beta^2 \sigma_{wds,V}^2(x_i, y_i)$, and $\sigma_{wdr,D}^2(i) = \alpha^2 \sigma_{wdr,L}^2(i) + \beta^2 \sigma_{wdr,V}^2(i)$. These equations provide a way in which the statistical model of gate delay (i.e., its three variances) can be computed from the underlying statistical model of transistor parameters.

5.2 Path Delay

Consider a path of N logic stages (gates). Variations in path delay are due to variations in the delays of both the gates and the interconnect. For simplicity of presentation, we will focus on the gate delay variations only. Interconnect delays can be handled in a similar way. Having assumed that nominal gate delay coincides with mean gate delay, the same follows for paths. Let $D_N(j)$ denote the deviation of the delay of path j from its mean (nominal) value. Since $D_N(j) = \sum_{i=1}^{N} D(i)$, then:

$$D_N(j) = N D_{dd} + \sum_{i=1}^{N} D_{wds}(x_i, y_i) + \sum_{i=1}^{N} D_{wdr}(i) \tag{44}$$

The gates on a path exist at various different locations. We will make the simplifying assumption that as far as physical location on the die, for purposes of computing the within-die-systematic component, all gates on path j share the same "nominal" coordinates (x_j, y_j), so that:

$$D_N(j) = N D_{dd} + N D_{wds}(x_j, y_j) + \sum_{i=1}^{N} D_{wdr}(i) \tag{45}$$

This is motivated by the expectation that gates on a critical path should be nearby on the die, and differences between their position-dependent within-die-systematic variations should be minor. Based on the independence relations between the terms in the above, we have $\sigma_{dd,D_N}^2 = N^2 \sigma_{dd,D}^2$, $\sigma_{wds,D_N}^2(x_j, y_j) = N^2 \sigma_{wds,D}^2(x_j, y_j)$, and if $\hat{\sigma}_{wdr,D}^2(j)$ is the average value of $\sigma_{wdr,D}^2(i)$ over all gates on this path, then $\sigma_{wdr,D_N}^2(j) = \sum_{i=1}^{N} \sigma_{wdr,D}^2(i) = N \hat{\sigma}_{wdr,D}^2(j)$. As for $\hat{\sigma}_{wdr,D}^2(j)$, we may approximate it using the average value of $\sigma_{wdr,D}^2(i)$ over the whole die, which we denote by $\hat{\sigma}_{wdr,D}^2$, so that:

$$\sigma_{wdr,D_N}^2(j) \approx N \hat{\sigma}_{wdr,D}^2 \tag{46}$$

With this, we have a full statistical model of path delay, so that we can treat it as a "parameter" and we can talk about its yield, as was done for the generic parameter $X(i)$ in sections 4 and 4.3.

5.3 Timing Yield

The timing success of an integrated circuit depends on a number of factors, including max delay violations, min delay violations, clock skew violations, etc. In this work, we focus on max delay constraints and consider a circuit to "pass" the timing test if its longest (critical) path delays are below some threshold (our work can be extended to cover other timing concerns). We let N be the number of stages (gates) on a path that would be representative of these critical paths, and we consider that the chip contains a (typically large) number of *disjoint* (non-intersecting) critical paths of N stages (gates) each, so that our expression for the *chip timing yield* becomes:

$$Y(\tau) = \mathcal{P}\{D_N(j) \le \tau,\ \forall j\} \ge Y_0(\tau) \tag{47}$$

where $Y_0(\cdot)$ is the lower bound expression for yield found above, in sections 4 and 4.3. Since the paths being considered are disjoint, then any correlations between their delays would be due only to correlations in the process variations, and not to the sharing of circuit component. If $\mathcal{Y}$ is the desired yield, then the techniques of sections 4 and 4.3 effectively provide the inverse function to compute τ for any desired $\mathcal{Y}$:

$$\tau = Y_0^{-1}(\mathcal{Y}) \tag{48}$$

where $Y_0^{-1}(\cdot)$ depends on the variances of D_N, which can be computed using the expressions for path and gate variances given above, from underlying transistor level variances. This τ is the timing margin of an N-gate path, for the desired specified yield $\mathcal{Y}$. Therefore, in order to get the desired yield, the circuit should be designed to "pass" the timing constraints when $D_N(j) = \tau$, for all j. Therefore, we set $D(i) = \tau/N$, and based on (42) we require:

$$\frac{\tau}{N} = \alpha L(i) + \beta V(i), \quad \forall i \tag{49}$$

This gives the range of possible settings of $L(i)$ and $V(i)$ required to achieve a timing deviation of τ on N-gate paths. If the circuit "passes" under these conditions, then the desired yield would be achieved. To simplify matters, suppose we want the $L(i)$ and $V(i)$ settings to be the same multiple of their individual σ's, i.e., let:

$$\frac{L(i)}{\sigma_L(i)} = \frac{V(i)}{\sigma_V(i)} = \delta, \quad \forall i \tag{50}$$

Notice that this is feasible (and $\delta > 0$) because both α and β have the same sign (positive). This leads to:

$$\delta = \frac{Y_0^{-1}(\mathcal{Y})/N}{\alpha \sigma_L(i) + \beta \sigma_V(i)} \tag{51}$$

This δ effectively defines the "worst-case file" for which the circuit should be tested (simulated, or checked) for timing constraint violations, so as to guarantee that the timing yield is at least $\mathcal{Y}$.

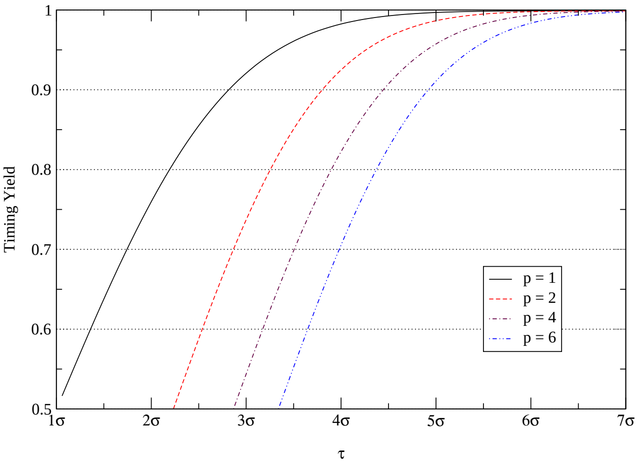

Figure 3: Timing yield plots, for $k = 3$ and $N = 9$.

6. APPLICATION TO STATIC TIMING

We will now illustrate how the above timing yield model allows us to choose a setting for the transistor parameters so that a desired yield is achieved if the circuit passes traditional static timing analysis. For clarity, let all transistor-level variances be equal: $\sigma^2_{dd,L} = \sigma^2_{wds,L}(x_i, y_i) = \sigma^2_{wdr,L}(i) = \sigma^2_L/3$, $\forall i$, and $\sigma^2_{dd,V} = \sigma^2_{wds,V}(x_i, y_i) = \sigma^2_{wdr,V}(i) = \sigma^2_V/3$, $\forall i$. At the gate level, this leads to $\sigma^2_{dd,D} = \left(\alpha^2\sigma^2_L + \beta^2\sigma^2_V\right)/3$, $\sigma^2_{wds,D}(x_i, y_i) = \left(\alpha^2\sigma^2_L + \beta^2\sigma^2_V\right)/3$, and $\sigma^2_{wdr,D}(i) = \left(\alpha^2\sigma^2_L + \beta^2\sigma^2_V\right)/3$. Therefore, $\sigma^2_D = \left(\alpha^2\sigma^2_L + \beta^2\sigma^2_V\right)$. At the path level, we have $\sigma^2_{dd,D_N} = \sigma^2_D N^2/3$, $\sigma^2_{wds,D_N}(x_j, y_j) = \sigma^2_D N^2/3$, and $\sigma^2_{wdr,D_N}(j) = \sigma^2_D N/3$, where the last equation is notable for the absence of the square factor. If we let $\sigma^2 = N^2\sigma^2_D/3$, then $\sigma_{dd,D_N} = \sigma_{wds,D_N} = \sigma$ and $\sigma_{wdr,D_N} = \sigma/\sqrt{N}$, where the last expression is notable for the presence of the square root term. With this, the equations for timing yield become as follows. In case $p = 1$, we have:

$$Y_0(\tau) = \int_0^1 \Phi\left((\tau/\sigma) - (k/\sqrt{N}) - \Phi^{-1}(u)\right) du \qquad (52)$$

and in the general case we have:

$$Y_0(\tau) = \int_0^{\Phi\left((\tau/\sigma)-(k/\sqrt{N})\right)} \chi^2_p\left(\left[(\tau/\sigma) - (k/\sqrt{N}) - \Phi^{-1}(u)\right]^2\right) du \qquad (53)$$

Plots of $Y_0(\tau)$ for a few values of p are shown in Fig. 3, for $k = 3$ and $N = 9$. One can use this type of figure as follows. If $p = 6$ and we want 90% yield (i.e., $\mathcal{Y} = 0.9$), then it is clear from the figure that we need $\tau \approx 5\sigma$, which leads to:

$$\delta = \left(\frac{5}{\sqrt{3}}\right)\frac{\sqrt{\alpha^2\sigma^2_L + \beta^2\sigma^2_V}}{\alpha\sigma_L + \beta\sigma_V} \qquad (54)$$

Let $r = (\alpha\sigma_L)/(\beta\sigma_V)$ then:

$$\delta = \left(\frac{5}{\sqrt{3}}\right)\frac{\sqrt{1 + r^2}}{1 + r} \qquad (55)$$

If, for example, $r = 1$, then:

$$\delta = \frac{5}{\sqrt{6}} \approx 2 \qquad (56)$$

so that the circuit would need to be simulated (and its timing checked) with all its transistors' L_e and V_t set at their $+2\sigma$ points. Notice that, since α and β depend on transistor sizing, then (55) provides a way in which δ can be controlled by circuit optimization and/or process tuning.

7. CONCLUSION

A method for statistical timing analysis has been developed, based on a timing yield model. The model is a "full-chip" model in that it can be applied with ease to large chips, before layout. This is achieved by using a measure of yield based on use of a *generic critical path* concept and capturing the statistics of a large collection of such paths using a model of within-die correlations that uses principal components analysis. This results in a methodology whereby one can select the right setting of the transistor parameters to be used in simulation or in traditional timing analysis in order to verify performance while guaranteeing a certain desired yield.

8. REFERENCES

[1] K. K. Low and S. W. Director. A new methodology for the design centering of IC fabrication processes. *IEEE Trans. on Computer-Aided Design*, 10(7):895–903, July 1991.

[2] R. W. Dutton and A. J. Strojwas. Perspectives on technology and technology-driven CAD. *IEEE Trans. on Computer-Aided Design*, 19(12):1544–1560, December 2000.

[3] M. Eisele, J. Berthold, D. Schmitt-Landsiedel, and R. Mahnkopf. The impact of intra-die device parameter variations on path delays and on the design for yield of low voltage digital circuits. *IEEE Trans. on VLSI*, 5(4):360–368, December 1997.

[4] A. Gattiker, S. Nassif, R. Dinakar, and C. Long. Timing yield estimation from static timing analysis. In *IEEE Int'l Symp. on Quality Electronic Design*, pages 437–442, San Jose, CA, March 26-28 2001.

[5] K. Singhal and V. Visvanathan. Statistical device models for worst case files and electrical test data. *IEEE Trans. on Semicon. Manufacturing*, 12(4):470–484, November 1999.

[6] L. Mizrukhin, J. Huey, and S. Mehta. Prediction of product yield distributions from wafer parametric measurements of CMOS circuits. *IEEE Trans. on Semiconductor Manufacturing*, 5(2):88–93, May 1992.

[7] C. Visweswariah. Death, taxes and failing chips. In *ACM/IEEE 40th Design Automation Conference*, pages 343–347, Anaheim, CA, June 2-6 2003.

[8] J. A. G. Jess, K. Kalafala, W. R. Naidu, R. H. J. M. Otten, and C. Visweswariah. Statistical timing for parametric yield prediction of digital integrated circuits. In *Design Automation Conf.*, pages 932–937, Anaheim, June 2-6 2003.

[9] A. Agarwal, D. Blaauw, V. Zolotov, and S. Vrudhula. Computation and refinement of statistical bounds on circuit delay. In *Design Automation Conf.*, pages 348–353, Anaheim, CA, June 2-6 2003.

[10] V. Mehrotra, S. L. Sam, D. Boning, A. Chandrakasan, R. Vallishayee, and S. Nassif. A methodology for modeling the effects of systematic within-die interconnect and device variations on circuit performance. In *Design Automation Conf.*, pages 172–175, Los Angeles, CA, June 5-9 2000.

[11] A. Devgan and C. Kashyap. Block-based static timing analysis with uncertainty. In *Int'l Conf. on Computer-Aided Design*, pages 607–614, San Jose, CA, November 9-13 2003.

[12] S. Bhardwaj, S. B.K. Vrudhula, and D. Blaauw. tau: Timing analysis under uncertainty. In *Int'l Conf. on Computer-Aided Design*, pages 615–620, San Jose, CA, November 9-13 2003.

[13] K. A. Bowman and J. D. Meindl. Impact of within-die parameter fluctuations on future maximum clock frequency distributions. In *Custom Integrated Circuits Conf.*, pages 229–232, 2001.

[14] H. Chang and S. S. Sapatnekar. Statistical timing analysis considering spatial correlations using a single PERT-like traversal. In *Int'l Conf. on Computer-Aided Design*, pages 621–625, San Jose, CA, November 9-13 2003.

[15] A. Agarwal, D. Blaauw, and V. Zolotov. Statistical timing analysis for intra-die process variations with spatial correlations. In *Int'l Conf. on Computer-Aided Design*, pages 900–907, San Jose, CA, November 9-13 2003.

[16] K. Okada, K. Yamaoka, and H. Onodera. A statistical gate-delay model considering intra-gate variability. In *Int'l Conf. on Computer-Aided Design*, pages 908–913, San Jose, CA, November 9-13 2003.

[17] A. Papoulis. *Probability, Random Variables, and Stochastic Processes.* McGraw-Hill, New York, NY, 2nd edition, 1984.

[18] S. G. Duvall. Statistical circuit modeling and optimization. In *Int'l Workshop on Statistical Metrology*, pages 56–63, June 2000.

[19] M. S. Srivastava. *Methods of Multivariate Statistics.* Wiley-Interscience, New York, NY, 2002.