Early Statistical Timing Analysis with Unknown Within-Die Correlations [*]

Khaled R. Heloue Farid N. Najm

Department of Electrical and Computer Engineering
University of Toronto, Toronto, Ontario, Canada
{khaled, najm}@eecg.toronto.edu

ABSTRACT

The ever increasing variability in process parameters, giving rise to circuit delay variations, presents an important challenge to the prediction and verification of circuit timing. On one hand, corner case analysis is rapidly becoming a daunting task, as exploring an exponentially increasing number of corners is practically infeasible. On the other hand, recently proposed approaches to statistical static timing analysis (SSTA) do not fit well in the timing verification methodology as they rely on the existence of correlation models that are not readily and realistically available from the process, at least not at an early stage of the design. In this paper, we present an early statistical timing analysis technique that can operate during pre-placement, when within-die correlations are still unknown. Starting from a simple delay model that requires minimal input from the user, we will predict bounds on the distribution of the maximum circuit delay. Such bounds are valid for any circuit placement, and consequently any arbitrary within-die correlation. We will use these bounds to introduce the concept of *margin uncertainty* and predict a margin range that can help designers at an early stage of the design flow.

1. INTRODUCTION

Process variations have an impact on circuit delay variations, and consequently can cause timing yield loss. Traditionally, process variations have been taken care of in various ways. In microprocessors, it is typical to check circuit timing with *nominal* transistor files, and to specify some *timing margin* that should be left as slack between the nominal delays and the timing constraints, in order to account for process variations. In ASICs, the practice is to typically design circuits by making sure the chip passes the timing requirements at all *process corners*, including nominal, worst, and best cases of device behavior. If these settings are too pessimistic, then designers are forced to waste time and effort optimizing a circuit using design conditions that are too stringent.

There has been considerable discussion in the literature that the traditional methods of using process corners or using a timing margin are breaking down. For microprocessors, where nominal process files are used and a timing margin needs to be left as slack, there is no easy way to decide what the margin should be, to account for within-die variations which have become important recently. And if a given margin is used, there is no indication as to what the resulting timing yield would be. On the other hand, for ASICs, and due to the increasing number of varying process parameters, the number of corners is exponentially increasing, thus making it very expensive to explore all corners.

Recently, statistical timing analysis techniques have been proposed as an alternative to corner case analysis. This is because corner analysis can not handle increasingly significant within-die variations [1]. Moreover, it is generally thought that traditional approaches are pessimistic and overly conservative. Since these statistical techniques aim to extend traditional static timing analysis (STA), they are better known as statistical static timing analysis (SSTA) techniques.

1.1 Types of SSTA

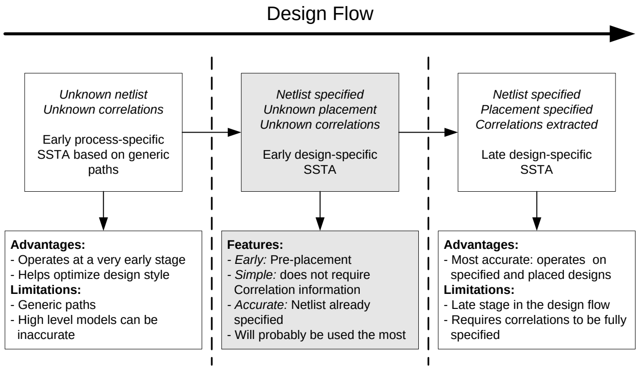
Design Flow
Unknown netlist
Unknown correlations
Early process-specific SSTA based on generic paths
Netlist specified
Unknown placement
Unknown correlations
Early design-specific SSTA
Netlist specified
Placement specified
Correlations extracted
Late design-specific SSTA
Advantages:
- Operates at a very early stage
- Helps optimize design style
Limitations:
- Generic paths
- High level models can be inaccurate
Features:
- Early: Pre-placement
- Simple: does not require Correlation information
- Accurate: Netlist already specified
- Will probably be used the most
Advantages:
- Most accurate: operates on specified and placed designs
Limitations:
- Late stage in the design flow
- Requires correlations to be fully specified

Figure 1: Types of SSTA

In the past few years, a wide variety of SSTA techniques have been proposed. These can be divided into two categories, namely path-based SSTA and block-based SSTA. Path-based techniques work by enumerating all paths and studying their joint distribution by integrating over the space of parameters [2]. In [3], a path-based technique which determines bounds on the distribution of circuit delay is proposed. This approach assumes that gate delay correlations are known beforehand, and therefore it can only be applied during post-placement. Other path-based approaches have been proposed [4] where within-die variations are assumed to be independent. Even though path-based techniques are known to be accurate, the exponential number of paths that need to be taken into account can be a major disadvantage.

Block-based techniques are more appealing. The two basic sum and max operations are performed on random variables, and propagation is done in a PERT-like fashion (topological order), which makes block-based approaches linear in the circuit size as long as the sum and max operations are performed in constant time.

In earlier work [5, 4, 6, 7], within-die variations are assumed to be totally uncorrelated. This assumption is not true in practice, however it is usually hard to express the correlations between within-die parameter variations with a model built from process data. Different attempts to model correlations have been proposed: In [8], principal component analysis (PCA) has been used to de-correlate variations on a set of independent random variables. In [9], a quad-tree partitioning is used to express a region-wise spatial correlation among within-die variations. In [10], correlation is taken care of using a canonical model, where each variation is expressed in terms of global sources of variations. All these methods depend on placement information, and rely on extensive process data, which is not readily available. Therefore, these types of post-placement SSTA become final sign-off tools and can not be used during early circuit design. In [11, 12], an early SSTA technique is presented. The analysis is based on a “generic path” concept that is representative of critical paths in a given technology.

Fig. 1 presents a general classification of SSTA techniques, and highlights where they fall in the design flow. In addition, it shows the advantages and limitations of each SSTA type in terms of chronology (early/late, pre-/post-placement) and accuracy. So far, only early process-specific (first type) and late post-placement design-specific (third type) SSTAs have been tackled in the literature. On one hand, the problem with the former type is the concept of generic path which might not capture well a specific design; on the other hand, the problem with the latter SSTA type is its reliance on extensive correlation and placement information, which may not be available. In this paper, we try to combine the best of the two worlds by presenting the first early pre-placement

[*] This project was supported in part by Intel Corp.

SSTA. As defined in Fig. 1, our technique is *early*, *simple*, and *accurate* since it operates pre-placement on a specific netlist without requiring any correlation information. This does not mean, however, that we ignore correlations; on the contrary, we handle the lack of correlation information using bounds which are valid for any arbitrary correlation or placement.

1.2 Yield Specific Margins

As was mentioned earlier, the goal of statistical timing analysis is to verify timing under process variations, and not to simply predict the distribution of circuit delay. The real question is *how much margin* should one leave on top of nominal maximum delay in order to guarantee a desired timing yield. SSTA answers this question by predicting the distribution of circuit delay. Assume that a target yield of 99% is desired; using SSTA, the 99^{th} delay percentile is predicted, from which the 99% yield margin is deduced by simply subtracting the nominal circuit delay. The following equation shows how to determine the timing margin for a yield of α:

$$\tau_\alpha = D_\alpha - D_{\text{nom}} \tag{1}$$

where D_{nom} is the nominal maximum circuit delay, D_α is the α-percentile of delay, and τ_α is the timing margin specific to yield α. Once this *yield specific margin* is determined, timing is verified by simply "reducing" the nominal circuit delay by this much margin.

In late post-placement SSTA, the design has already been placed and correlations have been extracted; therefore, a *unique* margin is predicted to verify the timing of the design. However, our technique falls into early pre-placement SSTA where correlations are still unknown; this uncertainty in the correlations is translated into a *margin uncertainty* as our technique will predict a margin *range* to cover all possible correlation settings. Fig. 4 gives a "sneak preview" of the big picture, where a *min margin* (predicted from the upper bound on delay distribution), a *max margin* (predicted from the lower bound on delay distribution), and a *margin uncertainty* (to account for correlations) are presented to designers at an early stage of the design flow; these margins can help them take early design decisions without having to wait until the design is placed. We also show that the margins predicted by post-placement SSTA for a specific correlation and placement fall within our margin range.

1.3 Overview

Starting from a simple process parameter model broken down into die-to-die and within-die components, and assuming a linear delay model based on sensitivities to process parameters, we will construct a standard delay model and propagate it in the timing graph similarly to what has been proposed in the literature [8, 10]. To account for the unknown within-die correlations however, we assess the effect of correlation on the sum and max of random variables to produce bounds on the distribution of circuit delay, and prove that these bounds are valid for arbitrary correlations. Using these bounds, we present the new concept of *margin uncertainty* that can help designers verify timing at an early stage of the design flow.

We believe that our approach would be a good addition to, and not a replacement for, the current post-placement SSTA techniques; our margin range can guide designers during early optimization, while the final margin predicted by post-placement SSTA can be used for final sign-off to achieve timing closure.

The rest of the paper is organized as follows: Section 2 presents the variational model, including process parameter model, gate delay, and arrival time model. Next, the effect of correlation on the sum and max of random variables is studied in section 3 where two important results are given. Section 4 gives a detailed overview of our analysis, which we validate in section 5. Finally, we propose an extension to our approach and conclude in sections 6 and 7.

2. MODELING VARIATIONS

In this section, we start by modeling variations at the level of process parameters. Then, using a first-order linear delay model, we express timing quantities such as gate delays and arrival times as a function of the underlying process variations.

2.1 Parameter Model

For a given circuit element or layout feature i, let $X_j(i)$, be a zero-mean Gaussian *random variable* (RV) that denotes the variation of a certain parameter j of this element from its nominal (mean) value. Thus, for example, $X_j(i)$ may represent channel length variations of transistor i. Notice, the Gaussian assumption is very common in the literature [2, 8, 10, 11, 12]. It is also standard practice [13] to express parameter variation by breaking it up into *die-to-die* and *within-die* components, as follows:

$$X_j(i) = X_{dd,j} + X_{wd,j}(i) \tag{2}$$

The die-to-die component $X_{dd,j}$ is an *independent*[1] zero-mean Gaussian RV that is *global* to the die as it takes the same value for all instances of this element on a given die, irrespective of location. The within-die component $X_{wd,j}(i)$ is a zero-mean Gaussian which can take different values for different instances of that element on the same die. It is thus a *local* variation specific to every instance. Keep in mind that $X_{wd,j}(i)$ has some correlation due to systematic effects. We can rewrite (2) in the following way:

$$X_j(i) = \sigma_{dd,j}\, Z_{dd,j} + \sigma_{wd,j}\, Z_{wd,j}(i) \tag{3}$$

where $\sigma_{dd,j}$ and $\sigma_{wd,j}$ are the parameter's die-to-die and within-die standard deviations respectively which can be obtained from the process for that specific parameter j. Note that $Z_{dd,j}$ and $Z_{wd,j}(i)$ are standard normal RVs with zero-mean and unit variance, and that $Z_{dd,j}$ is global whereas $Z_{wd,j}(i)$ is local with *some* correlation for different i's. For the scope of this work, the within-die correlation will be considered *unknown* or *unavailable.*

2.2 Gate Delay model

In general, there is a nonlinear relationship between gate delay and transistor parameters. Simple circuit simulations, however, reveal that this nonlinearity is not strong, especially for small transistor parameter variations. Therefore, we will simply assume that gate delay is linearly dependent on the process, and hence is Gaussian with mean equal to its nominal value. This assumption is also very common, and is used in all first-order path-based and block-based techniques [2, 8, 4, 3, 11].

Assume that p process parameters are varying; these can include channel length, threshold voltage, transistor width, and so on. For each gate, we can extract sensitivities to the different varying process parameters using circuit simulation; this can be done as part of library characterization. We can thus write the delay of gate i, $D(i)$, in the following way:

$$D(i) = \mu_i + \sum_{j=1}^{p} s_{ij} X_j(i) \tag{4}$$

where μ_i is the mean (nominal) delay, s_{ij} is the delay sensitivity of gate i to process parameter j, and $X_j(i)$ is the variation of process parameter j as defined in (3). Note that, for all the transistors within one logic gate i, we assume that the variations of process parameter j are captured with a *single* RV $X_j(i)$ and that $X_j(i)$'s are assumed to be independent for different j's, *i.e.*, variations of different process parameters are independent. Replacing $X_j(i)$ with its value from (3) yields:

$$D(i) = \mu_i + \sum_{j=1}^{p} \alpha_{ij} Z_{dd,j} + \sum_{j=1}^{p} \beta_{ij} Z_{wd,j}(i) \tag{5}$$

where $\alpha_{ij} = s_{ij}\,\sigma_{dd,j}$ and $\beta_{ij} = s_{ij}\,\sigma_{wd,j}$. We can further group the within-die variations of different parameters into a single within-die *delay* component. This leads to the following expression:

$$D(i) = \mu_i + \sum_{j=1}^{p} \alpha_{ij} Z_{dd,j} + \beta_{wd,i} Z_{wd,i} \tag{6}$$

[1] Throughout this paper, whenever an individual RV is described as "independent", this means that it is independent of all other RVs under consideration.

where $\beta_{wd,i} = \sqrt{\sum_{j=1}^{p} \beta_{ij}^2}$ since $Z_{wd,j}(i)$ are independent for different j. Note that $Z_{wd,i}$ is a standard normal RV that represents the within-die variation of delay $D(i)$. Also note that for different gates i and k, the within-die components of $D(i)$ and $D(k)$, *i.e.*, $Z_{wd,i}$ and $Z_{wd,k}$ will have some unknown correlation due to the correlation in process that is also considered unknown.

Generally, it is more accurate to specify a *timing arc delay* rather than a *gate delay*, since a gate can have different timing arcs, with different delays. Therefore, we will be expressing timing arc delays later on in the paper using the same delay model in (6).

2.3 Arrival Time model

Similarly, signal arrival times at the inputs and outputs of gates are modeled as normally distributed random variables using the same model for gate delays. Let A be a signal arrival time; then we can express A as follows:

$$A = a_o + \sum_{j=1}^{p} a_j Z_{dd,j} + a_{\mathrm{p}+1} Z_{wd,A} \tag{7}$$

where a_o is the mean of A, a_j's are the sensitivities to the (global) die-to-die components of the various process parameters, and $a_{\mathrm{p}+1}$ is the sensitivity to the (local) within-die component specific to A, *i.e.*, $Z_{wd,A}$. Note that similarly to gate delays, the within-die components of different arrival times will have some unknown correlations that particularly depend on placement and circuit topology.

Notice that unlike gate delay $D(i)$ in (6), where μ_i, α_{ij}, and $\beta_{wd,i}$ are inputs determined by the user through characterization, arrival time's parameters a_o, a_j's, and $a_{\mathrm{p}+1}$ will be determined by our algorithm through propagation in the timing graph. From here onwards, we will refer to our delay model as the **standard delay model**; this model has a constant part equal to the mean delay, a die-to-die part based on a linear expansion over the global die-to-die components of process parameters, and a within-die part that is local to the particular timing quantity in hand, which has some unknown correlation among different timing quantities.

3. EFFECT OF CORRELATION

As mentioned earlier, our approach does not require correlation information and is therefore valid before circuit placement. Having said this, we will look into ways to assess the effect of unknown correlation on the two timing operations that are used during propagation, *i.e.*, the **sum** and **max** operations.

3.1 Effect on the Sum of Two RVs

Let X and Y be two normally distributed random variables with means μ_X and μ_Y respectively, standard deviations σ_X and σ_Y respectively, and correlation coefficient ρ. Let $Z = X + Y$. Then Z is also normally distributed, with mean μ and variance σ^2 given by:

$$\begin{aligned} \mu &= \mu_X + \mu_Y && (8) \\ \sigma^2 &= \sigma_X^2 + \sigma_Y^2 + 2\sigma_{XY} && (9) \\ &= \sigma_X^2 + \sigma_Y^2 + 2\rho\sigma_X\sigma_Y && (10) \end{aligned}$$

where $\sigma_{XY} = \rho\sigma_X\sigma_Y$ is the covariance of X and Y.

This shows that the variance σ^2 of the sum of two random variables is an increasing function of their correlation coefficient ρ. In other words, if ρ is unknown, we are sure that the variance of the sum lies between a minimum achieved when $\rho = \rho_{\min} = 0$ (X and Y are uncorrelated), and a maximum achieved when $\rho = \rho_{\max} = 1$ (X and Y are totally correlated). Let $\sigma_{\min}^2$ and $\sigma_{\max}^2$ be these minimum and maximum variances, respectively. Then:

$$\begin{aligned} \sigma_{\min}^2 &= \sigma_X^2 + \sigma_Y^2 && (11) \\ \sigma_{\max}^2 &= (\sigma_X + \sigma_Y)^2 && (12) \end{aligned}$$

Since Z is a normally distributed RV, then we can write its distribution using the cumulative distribution function (cdf) of the standard normal, $\Phi(\cdot)$:

$$\mathcal{P}\{Z \le x\} = \Phi\left(\frac{x-\mu}{\sigma}\right) \tag{13}$$

Note that $\Phi(\cdot)$ is non-decreasing, therefore for $x - \mu \ge 0$, we can write the following bounds on the CDF of Z:

$$\Phi\left(\frac{x-\mu}{\sigma_{\max}}\right) \le \Phi\left(\frac{x-\mu}{\sigma}\right) \le \Phi\left(\frac{x-\mu}{\sigma_{\min}}\right) \tag{14}$$

An important result can be drawn from the above equation:

Result 1. *the cdf of the sum of two normal RVs with unknown correlation ρ (ranging from 0 to 1) can be bound by two extremes: setting $\rho = 0$ will lead to a minimum variance and thus an upper bound, while setting $\rho = 1$ will lead to a maximum variance and thus a lower bound on the distribution.*

Note that the bounds are valid only beyond the mean of the sum, *i.e.*, $x \ge \mu$, which is translated to be above the 50% line for normal distributions; this is obviously the more interesting yield range. This will be validated by our results.

3.2 Effect on the Max of Two RVs

Similarly, let X and Y be two jointly normally distributed RVs, with unknown correlation coefficient ρ. Let $Z = \max(X, Y)$. We are interested in assessing the effect of ρ on the distribution of Z. For this purpose, let $F_\rho(\cdot)$ be the **joint** distribution of X and Y, for a particular ρ. We can write the cdf of the max Z in the following way:

$$\begin{aligned} \mathcal{P}\{Z \le a\} &= \mathcal{P}\{\max(X, Y) \le a\} && (15) \\ &= \mathcal{P}\{X \le a\,,\, Y \le a\} && (16) \\ &= F_\rho(a) && (17) \end{aligned}$$

Therefore, the cdf of Z is equal to $F_\rho(a)$, and is thus correlation dependent. Using Slepian's theorem [14], we know that the joint distribution of two normal RVs X and Y, and consequently the distribution of their max Z, is an increasing function of their correlation coefficient ρ. Therefore we can draw our second important result:

Result 2. *The distribution of the max of two jointly normal RVs with unknown correlation ρ (ranging from 0 to 1) can be bound by two extremes; a lower bound on the distribution is achieved by setting $\rho = 0$, and an upper bound on the distribution is achieved by setting $\rho = 1$.*

4. TIMING ANALYSIS OPERATIONS

In this section, we present our statistical timing analysis technique and explain how it operates on a given circuit. To do that, we will first define its operation on a single gate, and then show how to repeatedly apply it in a block-based fashion on the whole timing graph.

Fig. 2 shows a gate with two inputs A and B and output C. The arrival times at inputs A and B are assumed to be known from previous stages, and are given in the standard delay form. Also assume that D_1 and D_2 are gate delay arcs for inputs A and B respectively, and are also given in the standard delay form. The arrival time at C will be equal to:

$$C = \max\left[(A + D_1)\,,(B + D_2)\right] \tag{18}$$

Thus, in order to determine the arrival time at the output of any gate, we need to perform two basic timing operations: a **sum** operation performed on a gate delay arc and an input arrival time, and a **max** operation performed on the two timing quantities resulting from the *additions*.

Since the correlation between the within-die variations of different timing quantities is unknown, we will use the results from section 3 to derive bounds on the distribution of C. Applying our approach on every gate, and propagating these bounds in the timing graph will lead to bounds on the maximum circuit delay.

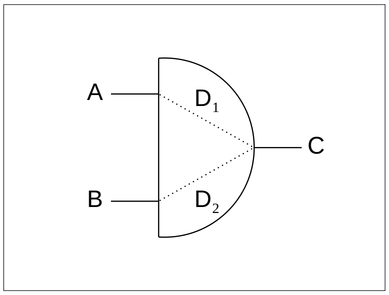

Figure 2: Single gate

4.1 Sum Operation

We will first show how to handle the sum operation given the lack of within-die correlation information. Assume that we have the gate in Fig. 2. Let $X = A + D_1$, and recall that both A and D_1 are expressed in the standard delay model:

$$\begin{aligned} A &= a_o + \sum_{j=1}^{p} a_j Z_{dd,j} + a_{\mathrm{p}+1} Z_{wd,A} && (19) \\ D_1 &= d_o + \sum_{j=1}^{p} d_j Z_{dd,j} + d_{\mathrm{p}+1} Z_{wd,D_1} && (20) \end{aligned}$$

We are interested in writing X using the standard model for variation in order to be able to propagate it to later stages in the timing graph. Simple addition leads to,

$$X = (a_o + d_o) + \sum_{j=1}^{p} (a_j + d_j) Z_{dd,j} + [a_{\mathrm{p}+1} Z_{wd,A} + d_{\mathrm{p}+1} Z_{wd,D_1}] \tag{21}$$

The above equation is not in standard delay form since the expression in brackets is not yet resolved as required. To do that, let $W = a_{\mathrm{p}+1} Z_{wd,A} + d_{\mathrm{p}+1} Z_{wd,D_1}$, and recall that the correlation between $Z_{wd,A}$ and Z_{wd,D_1} is unknown. Using the result of section 3.1, we want to bound the distribution of W and consequently the distribution of X. Let $Z_{wd,X_{lb}}$ be a standard normal RV. Then the variance of $(a_{\mathrm{p}+1}+d_{\mathrm{p}+1}) Z_{wd,X_{lb}}$ is $(a_{\mathrm{p}+1}+d_{\mathrm{p}+1})^2$ and, therefore, the distribution of $(a_{\mathrm{p}+1}+d_{\mathrm{p}+1}) Z_{wd,X_{lb}}$ bounds the distribution of W from below via $\sigma^2_{\max}$ as in (12) and (14). Similarly, let $Z_{wd,X_{ub}}$ be a standard normal RV. Then the variance of $\sqrt{a^2_{\mathrm{p}+1}+d^2_{\mathrm{p}+1}} Z_{wd,X_{ub}}$ is $(a^2_{\mathrm{p}+1}+d^2_{\mathrm{p}+1})$ and, therefore, the distribution of $\sqrt{a^2_{\mathrm{p}+1}+d^2_{\mathrm{p}+1}} Z_{wd,X_{ub}}$ bounds the distribution of W from above via $\sigma^2_{\min}$ as in (11) and (14). Now that we have derived bounds on W, we use them to bound the distribution of X. This is done by replacing the term in brackets, *i.e.*, W, in (21) by the two bounds. This gives us the two bounds on X:

$$X_{lb} = (a_o + d_o) + \sum_{j=1}^{p} (a_j + d_j) Z_{dd,j} + (a_{\mathrm{p}+1} + d_{\mathrm{p}+1}) Z_{wd,X_{lb}} \tag{22}$$

$$X_{ub} = (a_o + d_o) + \sum_{j=1}^{p} (a_j + d_j) Z_{dd,j} + \sqrt{a^2_{\mathrm{p}+1} + d^2_{\mathrm{p}+1}} Z_{wd,X_{ub}} \tag{23}$$

where X_{lb} and X_{ub} are two RVs whose distributions bound the distribution of X from below (lower-bound) and from above (upper-bound) respectively. Note that the within-die component of X_{lb} corresponds to the case where $Z_{wd,A}$ and Z_{wd,D_1} are assumed to be totally correlated ($\rho = 1$), which led to a maximum standard deviation of $(a_{\mathrm{p}+1}+d_{\mathrm{p}+1})$ as explained in section 3.1. Similarly, the within-die component of X_{ub} corresponds to the case where $Z_{wd,A}$ and Z_{wd,D_1} are assumed to be uncorrelated ($\rho = 0$), which led to a minimum standard deviation of $\sqrt{a^2_{\mathrm{p}+1}+d^2_{\mathrm{p}+1}}$.

Note that the within-die components of X_{ub} and X_{lb}, *i.e.*, $Z_{wd,X_{ub}}$ and $Z_{wd,X_{lb}}$ have lost any dependence on $Z_{wd,A}$ and Z_{wd,D_1}. This is not a problem, since our approach does not keep track of within-die correlation, but always considers it unknown. Hence, the correlation between $Z_{wd,X_{ub}}$ (or $Z_{wd,X_{lb}}$) and any other within-die component of other arrival times will be considered unknown. This is important, in order to account for any possible correlation.

Table 1: Within-die correlation settings for Sum and Max

	Lower Bound	Upper Bound
Sum	$\rho = 1$	$\rho = 0$
Max	$\rho = 0$	$\rho = 1$

At this point, we have handled the sum operation of a gate delay arc and an arrival time using bounds. For sake of clarity, we will refer to the process of generating a lower bound on the distribution of the sum, which was described above as **LB-sum**. Similarly, we will refer to the process of generating an upper bound on the distribution of the sum as **UB-sum**. The next section will show how we perform a max operation while keeping the standard delay model and handling unknown correlations.

4.2 Max Operation

It is well known that the max operator is not linear, which means that the maximum of two normally distributed RVs is not necessarily normal. Nevertheless, analytical expressions for the mean and variance of the maximum of two normally distributed RVs were determined by Clark in [15]. These expressions were presented and used in [8, 10], and we will review them next. Let $Z = \max(X, Y)$, then:

$$\begin{aligned} \mu_z &= \mu_x T + \mu_y (1 - T) + \theta \phi \left(\frac{\mu_x - \mu_y}{\theta} \right) && (24) \\ \sigma^2_z &= (\mu^2_x + \sigma^2_x) T + (\mu^2_y + \sigma^2_y)(1 - T) \\ &+ (\mu_x + \mu_y) \theta \phi \left(\frac{\mu_x - \mu_y}{\theta} \right) - \mu^2_z && (25) \end{aligned}$$

where $\phi(\cdot)$ is the probability density function (pdf) of the standard normal, and θ and T are given by:

$$\begin{aligned} \theta &= \sqrt{\sigma^2_x + \sigma^2_y - 2\rho\sigma_x\sigma_y} && (26) \\ T &= \Phi \left(\frac{\mu_x - \mu_y}{\theta} \right) && (27) \end{aligned}$$

The main idea is that, given the **means**, **variances**, and **correlation** coefficient of two normally distributed RVs X and Y, the *exact* mean and variance of the max Z can be determined. The max is then approximated by a normal distribution with the same mean and variance. In this section, we will use a similar approach and show how we handle the unknown within-die correlation.

Looking back at the gate in Fig. 2, assume that at this point we have performed the sum operation $X = A + D_1$ and $Y = B + D_2$, and that we need to *max* the timing quantities resulting from these two additions. For this purpose, let $C = \max(X, Y)$; recall that X and Y are already in the standard delay form:

$$\begin{aligned} X &= x_o + \sum_{j=1}^{p} x_j Z_{dd,j} + x_{\mathrm{p}+1} Z_{wd,X} && (28) \\ Y &= y_o + \sum_{j=1}^{p} y_j Z_{dd,j} + y_{\mathrm{p}+1} Z_{wd,Y} && (29) \end{aligned}$$

where $Z_{wd,X}$ and $Z_{wd,Y}$ are the within-die components of X and Y with unknown correlation ρ. We will use the result of section 3.2 to derive bounds on the distribution of $C = \max(X, Y)$. Let C_{ub} and C_{lb} be two RVs defined as follows:

$$\begin{aligned} C_{ub} &= \max(X, Y) \mid \rho_{X,Y} = \rho_{\max} && (30) \\ C_{lb} &= \max(X, Y) \mid \rho_{X,Y} = \rho_{\min} && (31) \end{aligned}$$

where $\rho_{X,Y}$ is the correlation coefficient between X and Y, and $\rho_{\max}$ ($\rho_{\min}$) is the maximum (minimum) possible correlation achieved by $\rho_{X,Y}$. Using the result from section 3.2, which states that the distribution of the max of two RVs is an increasing function of their correlation coefficient, it is easy to see that the distributions of C_{ub} and C_{lb}, as defined above, will bound the distribution of C from above and below, respectively.

To determine $\rho_{\max}$ and $\rho_{\min}$, we first express $\rho_{X,Y}$ in terms of the sensitivities of X and Y:

$$\rho_{X,Y} = \frac{\sum_{j=1}^{p} x_j y_j + \rho\, x_{\mathrm{p}+1} y_{\mathrm{p}+1}}{\sqrt{\sum_{j=1}^{p} x_j^2 + x_{\mathrm{p}+1}^2}\sqrt{\sum_{j=1}^{p} y_j^2 + y_{\mathrm{p}+1}^2}} \tag{32}$$

Note that $\rho_{X,Y}$ is also a function of the unknown within-die correlation ρ, and it can be shown that $\rho_{X,Y}$ is maximum when $\rho = 1$ and is minimum when $\rho = 0$. Therefore we can rewrite (30) and (31) in the following way:

$$C_{ub} = \max(X, Y) \mid \rho = 1 \tag{33}$$

$$C_{lb} = \max(X, Y) \mid \rho = 0 \tag{34}$$

Now that we have defined C_{ub} and C_{lb}, we express these timing quantities in the standard delay model to allow further propagation in the timing graph. To do this, we first use Clark's expressions to determine the exact mean and variance of C_{ub} (with $\rho = 1$) and C_{lb} (with $\rho = 0$). Then, we use the approach of [8, 10] to expand C_{ub} and C_{lb} in the standard delay model. This is done by preserving the dependence on the *global* RVs, *i.e.*, $Z_{dd,j}$, and then computing the within-die part by matching the total variance to the exact variance determined from Clark's expressions. This approach has become the standard way to approximate the max operation while preserving the correlation due to the global die-to-die variables. In the following, we will refer to the process of generating an upper/lower bound on the distribution of the max, *i.e.*, C_{ub}/C_{lb} as **UB/LB-max**.

4.3 Combining Sum and Max

In sections 4.1 and 4.2, we presented ways to find upper and lower bounds on the distributions of the sum and max of two RVs that are represented in our standard delay form. Recall that we are interested in deriving bounds on the distribution of C, the arrival time at the output of a gate. We know that:

$$C = \max\left[(A + D_1), (B + D_2)\right] \tag{35}$$

so that C can be determined by a series of two additions, $A + D_1$ and $B + D_2$, and one max involving the results of these additions. Therefore, in order to determine a lower bound on the distribution of C, we need to perform the sum and the max in the *lower bound mode*, *i.e.*, using LB-sum and LB-max. Similarly, to determine an upper bound on the distribution of C, we need to perform our analysis in the *upper bound mode*, *i.e.*, using UB-sum and UB-max successively.

Table 1 gives a summary of the different settings of the unknown within-die correlation ρ that should be used for the sum and max operations to get a lower/upper bound on the distribution of delay at the output node of a given gate.

4.4 Block-based Propagation

The rest of our approach is similar to deterministic STA, except: the deterministic sum is replaced by the UB and LB sums as described in section 4.1 and the deterministic max is replaced by the UB and LB max as described in section 4.2. In this way, and only in **a single pass**, we produce an upper and a lower bound on the distribution of delay at every node in the timing graph, particularly at the sink node (the circuit primary output). We are specifically interested in the delay of the sink node since it is equal to the maximum circuit delay.

Recall that the bounds produced are valid for any within-die correlation and consequently any circuit placement, since the analysis was performed under unknown correlation. This allows designers to be at an advantage, as one is able to predict, at an early stage of the design flow, the extent of circuit delay variations without having access to placement information.

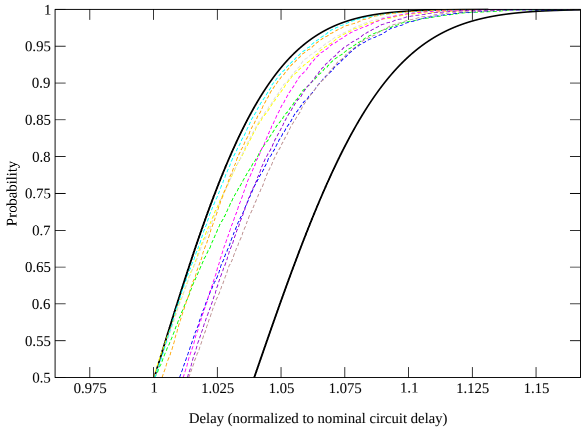

Figure 3: Upper/Lower bounds vs. MC distributions for circuit c499

5. RESULTS

To test our SSTA technique on realistic circuits, we have implemented it as part of an existing static timing analyzer using C/C++. A 90nm CMOS library of gates was characterized using HSPICE in order to determine the sensitivities of gate delay to various process parameters. We have chosen to vary channel length (L_n and L_p), and threshold voltage (V_{tn} and V_{tp}) of n-mos and p-mos transistors, and have determined the sensitivities of gate delays to these process parameters. Also, in all experiments, we have assumed equal within-die and die-to-die process variation, *i.e.*, the within-die and the die-to-die variance of each process parameter is equal to 50% of the total parameter variance:

$$\sigma_{dd}^2 = \sigma_{wd}^2 = \frac{1}{2}\sigma^2 \tag{36}$$

We ran our SSTA analyzer on all of the ISCAS85 benchmark circuits, and computed bounds on the distribution of the maximum circuit delay. In order to validate these bounds, we performed several Monte Carlo (MC) tests under arbitrary within-die correlations, generated using the grid model [8] for spatial correlation. We varied the number of grids to increase/decrease the correlation, and also generated cases of *biased* correlations where we increased correlations across some paths (to maximize variance of delay) or decreased correlations across primary inputs (to increase the mean of delay).

The results for circuit c499 are shown in figure 3, where the solid curves represent the upper and lower bounds on the distribution of delay, and the dotted curves represent the distributions of delay generated through MC analysis under different arbitrary and biased correlations. It is clear that all MC generated distributions fall between our bounds. Note that for higher percentiles, the bounds get tighter. Also note that, as stated at the end of section 3.1, our bounds are only valid above the 50% line. This is not a problem, since we are mainly interested in high delay percentiles in order to design for a high yield. The rest of the ISCAS circuits show similar results but are not shown for lack of space.

As was mentioned in section 1.2, the purpose of SSTA is to predict a *yield specific timing margin*, that is, the margin that should be left on top of the nominal maximum circuit delay in order to meet a certain target yield. Using the bounds on the circuit delay distribution, we can predict a min margin (from upper bound) and a max margin (from lower bound) by subtracting the nominal circuit delay from the desired delay percentiles predicted by the bounds as was explained in (1). Fig. 4 represents a plot to scale of these margins at a 99% target yield for all ISCAS circuits. It also shows the margin uncertainty caused by the unknown correlations that needed to be accounted for. Note that in order to guarantee the target yield, one needs to use the max margin since it is the one predicted from the lower bound on the distribution of delay, which accounts for "worst" possible correlations. The min margin is also useful to check if the design is feasible;

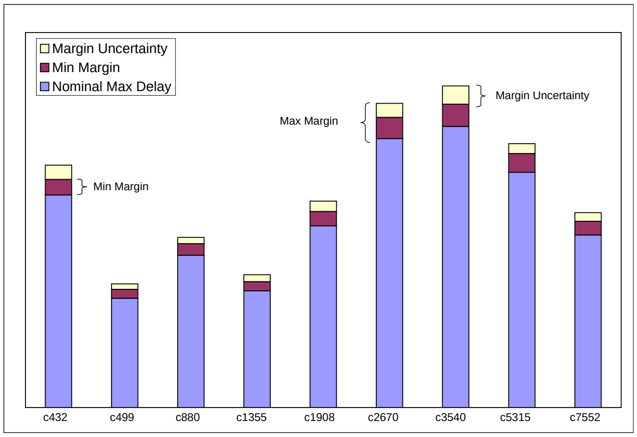

Figure 4: 99% yield Margins

Table 2: Margin Comparison as % of nominal max delay

ISCAS 85 **Circuit**	**Max Margin at** 99% **yield**	**"worst case" Margin**
c432	14.0%	24.0%
c499	13.2%	27.0%
c880	11.7%	24.4%
c1355	13.8%	24.2%
c1908	13.6%	25.4%
c2670	13.1%	26.6%
c3540	14.4%	26.4%
c5315	12.2%	26.1%
c7552	13.0%	25.3%

recall that since it accounts for "best" possible correlation, it is the smallest possible margin for the current design. If the min margin turned out to be unacceptable (too large) for designers, then the whole design needs to be adjusted. One conclusion to draw here is that, in the absence of any correlation information, and at a stage where the circuit has not yet been placed, it is a powerful asset to have an idea about the range of margins needed to achieve a desired yield. This allows designers to take early design decisions, and makes room for early optimization.

Table 2 lists, for all ISCAS circuits, the max margin as a percentage of nominal maximum circuit delay, at a 99% target yield. The largest max margin is about 14%. We have also listed, for sake of comparison, the margin predicted from "worst case" analysis on every gate. What we mean by "worst case" analysis is to set the within-die and die-to-die variations of every process parameter to their maximum (3σ) and perform deterministic STA. It is clear that the "worst case" margin is more pessimistic which shows the need for statistical techniques.

6. "INCOMPLETE CORRELATION INFORMATION"

The analysis presented so far assumes *unknown* within-die correlations, and hence tries to account for this lack of information by selecting extreme correlation values for each timing operation as specified in Table 1. Under such a *blind* setting, *i.e.*, completely unknown correlations, the extremes are clearly $[0, 1]$. However, it is not unusual that designers have *some* raw information about within-die correlations, either through access to floorplanning information, or through pure historical observations; these cases will be referred to as "cases of incomplete correlation information". We are mainly interested in a lowest (ρ_{min}) and highest (ρ_{max}) *observed* correlations, which, if available, can be safely used in lieu of the current pair of extremes $[0, 1]$, and the rest of the analysis remains the same. This can be viewed as an extension to our work, and we are currently investigating what a reasonable form of "incomplete correlation information" would be, from which we can draw ρ_{min} and ρ_{max}.

7. CONCLUSION

In this work, we have a presented a pre-placement statistical timing analysis technique that can operate under unknown within-die correlations. Starting from a simple delay model based on sensitivities to process parameters, we construct a standard delay model for arrival times, and propagate this model in the timing graph in a block-based fashion. To account for unknown correlations, we use bounds on each timing operation (sum and max) by taking the correlation coefficient to its extremes. We prove that these bounds are valid for any within-die correlation and thus any placement, which makes our analysis valid during pre-placement. We believe that our technique is powerful because it is able to predict the range of margins needed to achieve a target yield at a very early stage of the design flow, prior to the access to correlation information.

8. REFERENCES

[1] P. Zuchowski, P. Habitz, J. Hayes, and J. Oppold. Process and environmental variation impacts on asic timing. In *IEEE/ACM International Conference on Computer Aided Design*, pages 336–342, San Jose, CA, November 7-11 2004.

[2] 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, CA, June 2-6 2003.

[3] M. Orshansky and A. Bandyopadhyay. Fast statistical timing analysis handling arbitrary delay correlations. In *Design Automation Conference*, pages 337–342, San Diego, CA, June 7-11 2004.

[4] 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.

[5] 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.

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

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

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

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

[10] C. Visweswariah, K. Ravindran, K. Kalafala, S. Walker, and S. Narayan. First-order incremental block-based statistical timing analysis. In *Design Automation Conference*, pages 331–336, San Diego, CA, June 7-11 2004.

[11] F. Najm and N. Menezes. Statistical timing analysis based on a timing yield model. In *Design Automation Conference*, pages 460–465, San Diego, CA, June 7-11 2004.

[12] K. R. Heloue and F. N. Najm. Statistical timing analysis with two-sided constraints. In *IEEE/ACM International Conference on Computer Aided Design*, pages 829–836, San Jose, CA, November 6-10 2005.

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

[14] Y. L. Tong. *The multivariate normal distribution*. Springer-Verlag, New York, NY, 1990.

[15] C. E. Clark. The greatest of a finite set of random variables. In *Operaions Research*, pages 145–162, March-April 1961.