Statistical Timing Analysis for Intra-Die Process Variations with Spatial Correlations

Aseem Agarwal, David Blaauw, *Vladimir Zolotov

University of Michigan, Ann Arbor, MI
*Motorola, Inc., Austin, TX

Abstract

Process variations have become a critical issue in performance verification of high-performance designs. We present a new, statistical timing analysis method that accounts for inter- and intra-die process variations and their spatial correlations. Since statistical timing analysis has an exponential run time complexity, we propose a method whereby a statistical bound on the probability distribution function of the exact circuit delay is computed with linear run time. First, we develop a model for representing inter- and intra-die variations and their spatial correlations. Using this model, we then show how gate delays and arrival times can be represented as a sum of components, such that the correlation information between arrival times and gate delays is preserved. We then show how arrival times are propagated and merged in the circuit to obtain an arrival time distribution that is an upper bound on the distribution of the exact circuit delay. We prove the correctness of the bound and also show how the bound can be improved by propagating multiple arrival times. The proposed algorithms were implemented and tested on a set of benchmark circuits under several process variation scenarios. The results were compared with Monte Carlo simulation and show an accuracy of 3.32% on average over all test cases.

1 Introduction

Static timing analysis has become an indispensable part of performance verification. Static timing analysis has the advantage that it does not require input vectors and has a run time that is linear with the size of the circuit. A number of methods have been proposed to increase the accuracy of static timing analysis through improved delay models and analysis techniques. In recent technologies, the variability of circuit delay due to process variations has become a significant concern. As process geometries continue to shrink, the ability to control critical device parameters is becoming increasingly difficult, and significant variations in device length, doping concentrations, and oxide thicknesses have resulted.

Traditionally, process variations have been modeled in static timing analysis (STA) using so-called case analysis. In this methodology, best-case, nominal and worst-case SPICE parameters sets are constructed and the timing analysis is performed several times, each time using one case file. Each execution of static timing analysis is therefore deterministic, meaning that the analysis uses deterministic delays for the gates and any statistical variation in the underlying silicon is hidden. While this approach has been successfully used in the past to model die-to-die variations, it is not able to accurately model variations within a single die. With the continual scaling of feature sizes, the ability to control critical device parameters on a single die has become increasingly difficult. Using a worst-case analysis for these so-called intra-die variations therefore leads to very pessimistic analysis results since it assumes that all devices on a die have worst-case characteristics, ignoring their inherent statistical variation. The emerging dominance of intra-die variations therefore poses a major obstacle for deterministic STA, giving rise to the need for *statistical* timing analysis approaches.

ICCAD '03, November 11-13, 2003, San Jose, California, USA.

In general, process variations can be divided into *inter-die* variations and *intra-die* variations. Inter-die variations are variations that occur from one die to the next, meaning that the same device on a chip has different features among different die of a wafer, from wafer to wafer, and from wafer lot to wafer lot. Intra-die variations are variations in device features that are present within a single chip, meaning that a device feature varies between different locations on the same die. Intra-die variation result from equipment limitations or statistical effects in the fabrication process, such as statistical variations in the doping concentrations.

Intra-die variations often exhibit spatial correlations, where devices that are close to each other have a higher probability of being alike than devices that are placed far apart. This has been reported especially for gate length variations [1]. Intra-die variations can also have a deterministic component due to topologically dependencies of device processing, such as CMP effects and optical proximity effects [2]. In some cases, such topological dependencies can be directly accounted for in the analysis [3][4], whereas in other cases, such variations are treated as random.

Statistical timing analysis is similar to deterministic timing analysis in that arrival times are propagated through the circuit from primary inputs to primary output. In statistical timing analysis, however, the gate delays and arrival times are represented with random variables. The difficulty of statistical timing analysis results from the correlations that arise among the arrival times in the circuit and between the arrival times and gate delays. These correlations must be taken into account when arrival times are propagated in the circuit, leading to an exponential run time complexity and making statistical timing analysis a challenging problem.

A number of statistical timing analysis approaches have been proposed in recent years [5-18]. In [13] the correspondence between deterministic timing analysis and statistical timing analysis was first shown. However, the proposed method does not address the correlation between the arrival times. In [14], a novel method using discretized probability distributions is proposed. However, the run time of the method is exponential and the proposed approaches to reduce the run time have an unclear impact on the accuracy. In [15], a novel method using statistical bounds is proposed with gate delays restricted to Gaussian distributions. However, to obtain a high quality bound, it is necessary to enumerate all paths in the circuit, leading to exponential run time complexity. In [16], a path based statistical delay computation is presented using an accurate delay model. However, the analysis is performed one path at a time and the number of critical and near-critical paths in a circuit can be very large. In [17], a new circuit optimization method was therefore proposed that reduces the number of near critical paths in a circuit, thereby improving the statistical delay of the circuit. Finally, in [18] a method using statistical bounds is presented that addresses the arrival time correlations due to path reconvergence. However, the method does not address arrival time correlations due to spatial correlations between the gate delays.

In this paper, we therefore propose a new statistical timing analysis approach to model the impact of process variations on circuit

delay. We model both inter- and intra-die process variations and account for spatial correlations of the gate delays. In our analysis, we focus on gate length variability since it has been shown to have a dominant impact on gate delay [1]. However, our analysis can be easily extended to other process variations as well. We first present a model for inter- and intra-die gate length variation and their spatial correlations. Gate delays and arrival times are represented as a sum of random variables, and preserve the spatial correlation information.

The correlation between the arrival times complicates the computation of the maximum arrivals times, as required during arrival time propagation. Since the exact computation of the maximum arrival time requires exponential run time, we propose a method that produces an upper bound on the exact arrival time in linear run time. We prove the correctness of the proposed bound in the presence of spatially correlated gate delays. The obtained bound is itself a random variable with a probability distribution function, allowing for the computation of useful statistical quantities such as confidence points. In order to improve the proposed bound, we propose a method whereby multiple arrival times are propagated in the circuit at the expense of additional run time. We implemented the proposed methods and tested them on benchmark circuits. We demonstrate that using the proposed methods, the statistical delay of a circuit can be computed with high accuracy.

The remainder of this paper is organized as follows. In Section 2, we present our model of process variations and our modeling assumptions. In Section 3 we present our approach for statistical timing analysis. In Section 4, we present the heuristic method for improving the quality of the bound by propagating multiple arrival times. In Section 5, we present our results and in Section 6 we draw our conclusions.

2 Process Variation Model

In this section, we present our model for process variations. We consider two basic types of process variations in our analysis: inter-die variations and intra-die variations. Intra-die variation can be further divided into random variations, and spatially correlated variations. Random intra-die variations have no dependence on the location of the devices, while intra-die variations that are spatially correlated produce an increased likelihood of similar gate lengths for devices that are closely spaced versus those that are placed further apart. We first discuss our model for inter- and intra-die variations which is based on the model in [19] and then discuss how this model is extended to account for spatial correlations.

We propose the following model, where the device length $L_{total,k}$ of device k is the algebraic sum of the nominal gate length, the inter-die device length variation ΔL_{inter} and intra-die device length variation, $\Delta L_{intra,k}$:

$$L_{total,k} = L_{nom} + \Delta L_{inter} + \Delta L_{intra,k}, \qquad \text{(EQ 1)}$$

where ΔL_{inter} and $\Delta L_{intra,k}$ are random variables. L_{nom} represents the mean of the gate length across all possible die. All devices on a die share one variable ΔL_{inter} for the inter-die component of their total device length variation, which represents a variation of the *chip mean* of the gates of a particular die. $\Delta L_{intra,k}$ represents the variation of an individual gate from this *chip mean*. For the moment, we ignore the spatial correlation of intra-die variations, and hence each device is represented with a separate independent random variable $\Delta L_{intra,k}$, where all random variables $\Delta L_{intra,k}$ have identical probability distributions. For the purpose of our discussion, we assume that both random variables ΔL_{inter} and $\Delta L_{intra,k}$ have a truncated normal distribution. This reflects the fact that the gate length in an operational chip cannot be less than some finite minimum value or more than some finite maximum value. However, any suitable distribution can be used, and our proposed approach is not restricted to normal distributions.

After defining a model for the gate length variation, the delay d_k of gate k is now defined as follows:

$$d_k = D_k(L_{nom} + \Delta L_{inter} + \Delta L_{intra,k}) \qquad \text{(EQ 2)}$$

Since function D_k is in general a non-linear function, finding the distribution of d_k can be difficult. However, we take advantage of the fact that the gate length variations ΔL_{inter} and $\Delta L_{intra,k}$ are typically small, with typical 3-sigma values of less than 15% of L_{nom}. Hence, we make the simplifying assumption that, for small variations, the change in gate delay is linear with the change in gate length. Hence, we can write EQ2 as follows:

$$d_k = D_k(L_{nom}) + \Delta D_k(\Delta L_{inter}) + \Delta D_k(\Delta L_{intra,k}), \qquad \text{(EQ 3)}$$

where $\Delta D_k(L_{inter})$ and $\Delta D_k(L_{intra,k})$ are the change of gate delay due to inter- and intra-die gate length variation. For convenience, we define $\Delta D_k()$ as follows:

$$\Delta D_k(\Delta L) = \frac{\partial D_k}{\partial L} \times \Delta L, \qquad \text{(EQ 4)}$$

where the sensitivity of the delay with respect to device length $\partial D_k/\partial L$ is computed at the nominal device length. We can now express the delay of a gate with the following simple expression:

$$d_k = D_{nom} + \alpha\Delta L_{inter} + \alpha \Delta L_{intra,k} \qquad \text{(EQ 5)}$$

where $\alpha = \partial D_k/\partial L$. Note that instead of using EQ4 any linear fitting function could be used as well. Although EQ5 uses a simple linear approximation, such an approximation was found to give very good accuracy for current process variabilities [16][19].

Spatial Correlation Model

In EQ1, the intra-die variation of gate delay is modeled by assigning an independent random variable for each gate. However, in the presence of spatial correlation, these random variables become dependent which greatly complicates the analysis. We therefore propose the following method for modeling spatial correlation of intra-die process variation.

We first divide the area of the die into regions using a multi-level quad-tree partitioning, as shown in Figure 1. For each level l, the die area is partitioned into 2^l-by-2^l squares, where the first or top level 0 has a single region for the entire die and the last or bottom level m has 4^m regions. We then associate an independent random variable $\Delta L_{l,r}$ with each region (l, r) to represent a component of the total intra-die device length variation. The variation of a gate k is then composed as the sum of intra-die device length components $\Delta L_{l,r}$ where level l ranges from 0 to m and the region r at any particular level is the region that intersects with the position of gate k. For the gate in region 2,1 in Figure 1, the components of intra-die device length variation are therefore $\Delta L_{0,1}$, $\Delta L_{1,1}$ and $\Delta L_{2,1}$. The intra-die device length of gate k is now defined as the sum of all random variables $\Delta L_{l,r}$ associated with a gate:

$$\Delta L_{intra,k} = \sum_{0<l<m,\ r \text{ intersects } k} \Delta L_{l,r} + \Delta L_{random,k}, \qquad \text{(EQ 6)}$$

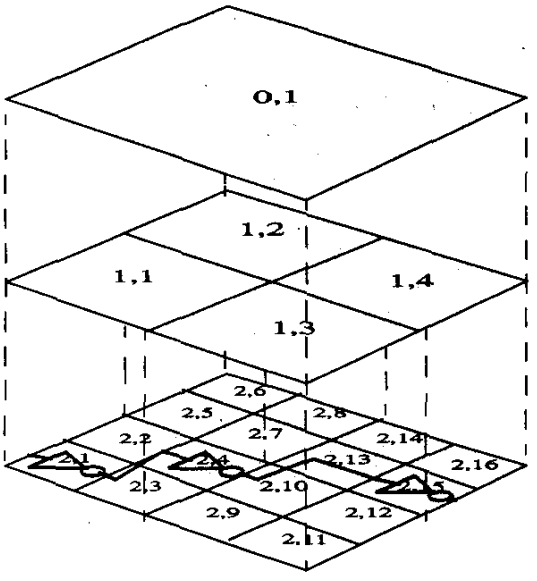

Figure 1. Modeling spatial correlations using quad-tree partitioning

where $\Delta L_{l,r}$ are the random variables associated with the quad-tree and $\Delta L_{random,k}$ is an independent random variable, assigned to each gate to model uncorrelated delay variation.

It must be ensured that the sum of all random variables $\Delta L_{l,r}$ associated with a gate always adds up to the total intra-die gate length variation. This can be accomplished by assigning all random variables associated with a particular level the same probability distribution and by dividing the total intra-die variability among the different levels.

Using the described model, gates that lie within close proximity of each other will have many common intra-die device length components resulting in a strong intra-die length correlation. Gates that lie far apart on a die share few common components and therefore have weak correlation. For the three gates shown in Figure 1 in regions (2,1), (2,4) and (2,15) the intra-die device length variation is expressed as follows:

$$\Delta L_{intra,1} = \Delta L_{2,1} + \Delta L_{1,1} + \Delta L_{0,1} + \Delta L_{random,1} \quad \text{(EQ 7)}$$

$$\Delta L_{intra,2} = \Delta L_{2,4} + \Delta L_{1,1} + \Delta L_{0,1} + \Delta L_{random,2} \quad \text{(EQ 8)}$$

$$\Delta L_{intra,3} = \Delta L_{2,15} + \Delta L_{1,4} + \Delta L_{0,1} + \Delta L_{random,3} \quad \text{(EQ 9)}$$

We can observe from the above equations that gates 1 and 2 are strongly correlated, as they share the common variables $\Delta L_{1,1}$ and $\Delta L_{0,1}$. On the other hand, gates 1 and 3 are more weakly correlated as they share only the common variable $\Delta L_{0,1}$. Note that the devices that are closely spaced, but fall in different squares, will have less correlation than those that are equally spaced, but fall within the same square. However, this issue can be addressed by using an additional quad-tree which is offset by half the size of the smallest square.

Figure 1 shows an example of a die with 3 levels of partitioning resulting in 16 regions at the bottom level. Since the number of regions at the bottom level grows as 4^m it is possible to obtain a fine partitioning of the die with only a moderate number of levels. Note also that length $\Delta L_{0,1}$ associated with the region at the top level of the hierarchy is equivalent to the inter-die device length ΔL_{inter} since it is shared by all gates on the die.

We can control how quickly the spatial correlation diminishes as the separation between two gates increases by controlling the allocation of total intra-die device length variation among the different levels. If the total intra-die variance is largely allocated to the bottom levels, and the regions at top levels have only a small variance, there is less sharing of device length variation between gates that are far apart and the spatial correlation will diminish quickly. On the other hand, if the total intra-die variance is predominantly allocated to the regions at the top levels of the hierarchy, then even gates that are widely spaced apart will still have significant correlation and spatial correlation will diminish more slowly as spacing increases. The proposed model is therefore flexible and can be easily fit to measured device length data.

Based on the above model for intra-die spatial correlation, we can combine EQ5 and EQ6 to obtain the following expression of the delay a gate:

$$d_k = D_{nom} + \quad \text{. (EQ 10)}$$

$$\alpha \cdot \left(\Delta L_{inter} + \sum_{0 < l < m,\ r \text{ intersects k}} \Delta L_{l,r} + \Delta L_{random,k} \right)$$

Note that all random variables in EQ10 are *independent* random variables. This has the advantage that spatial correlations can be processed using only independent random variables, which simplifies the analysis. Note also that some of the random variables in EQ10 will occur in the expressions of multiple gate delays.

Finally, to simplify the notation, we rewrite EQ10 using a more general form as follows:

$$d_k = D_{nom} + \sum_i \alpha_i \cdot L_i + \Delta D_{random,k} \quad \text{(EQ 11)}$$

Where L_i and $\Delta D_{random,k}$ are random variables and α_i are constants. $\Delta D_{random,k}$ is the random delay due to uncorrelated intra-die gate length variation. The variables L_i correspond to one of the random variables in the proposed model, such as ΔL_{inter} and $\Delta L_{l,r}$. The sum is taken over all random variables present in the model and $\alpha_i = \alpha$ for the random variable ΔL_{inter} and for the random variables $\Delta L_{l,r}$ associated with the gate, based on its position in the die. For all other i, $\alpha_i = 0$. Note that EQ11 is simply a more general and convenient form of EQ10, where the delay of a gate is expressed in terms of all random variables in the model, instead of just those associated with that particular gate. Using EQ11, the delay of a gate is expressed as a sum of independent random variables, some of which may be shared in the delay expression of one or more gates. In the following Section, we show how to perform timing analysis based on the proposed model for process variation.

3 Statistical Timing Analysis Method

Static timing analysis is performed by propagating arrival times from the primary inputs to the primary outputs using repeated application of two operations:

1. **Propagation**. Arrival times are propagated from the input of a gate to the output of that gate. In the process, the delay of the gate is added to the arrival time.

2. **Merging**. Multiple arrival times that converge at a gate output from different gate inputs are merged by taking the maximum of these arrival times.

Statistical timing analysis can be performed in the same manner using propagation and merging, except that both the gate delays and the arrival times are now random variables. In this case, the arrival time is specified either with a cumulative distribution function (*CDF*) or probability density function (*PDF*). To simplify the implementation of statistical STA it is often more convenient to approximate continuous PDFs and CDFs with discrete functions. For computational efficiency, we use discrete PDFs and CDFs in the implementation of our proposed method. However, for generality, we will formulate the statistical timing analysis task using continuous functions.

The difficulty in statistical timing analysis arises from the correlations between the random variables, which arise from one of two sources. First, reconvergence of circuit paths results in arrival times that are dependent, since they share a common portion of their path delay. However, in [18] it was shown that ignoring the correlation resulting from reconvergent fanout produces an upper bound on the statistical delay and results in a conservative analysis.

The second source of dependence results from spatial correlations between gate delays. It is clear that if the delay of two gates is correlated, the arrival times at their outputs will be correlated as well thereby complicating the merging operation of these two arrival times. Furthermore, spatial correlation also results in dependence between an arrival time and the gate delays themselves. This complicates the propagation operation where the delay of a gate is added to the arrival time at its input node.

It is easy to show that, unlike correlations resulting from reconvergent paths in the circuit, ignoring spatial correlations may not result in an upper bound on the statistical delay. This is intuitively obvious from the fact that spatial correlation makes the intra-die variability more similar to that of inter-die variability, which increases the delay of circuit paths. The correlation between the arrival times and between arrival times and the gate delays must therefore be accounted for during the propagation and merging operation.

Note that if we express the delay of a gate using a single random variable, by convolving its independent components in EQ11, it will be very difficult to recover the correlation information between this gate delay and another. In the proposed approach, we therefore maintain the representation of the delay of a gate using its sum of components, as shown in the right hand side of EQ11. Similarly, we need to preserve the correlation information of arrival times. Hence, we also represent the arrival times in the timing analysis using a sum of components. Similar to that of the gate delay in EQ11, an arrival time a is therefore expressed as follows:

$$a = A_{nom} + \sum_i \beta_i \cdot L_i + \Delta A_{random} \qquad \text{(EQ 12)}$$

where A_{nom} is the arrival time at nominal process conditions, L_i are the random variables of gate length, β_i are constant coefficients and ΔA_{random} is the uncorrelated component of arrival time variation. We will show that by expressing the arrival times in the same form as that of gate delay, their correlations can be determined and correctly addressed.

Using the proposed representations for gate delays and arrival times, we now perform arrival time propagation and merging, such that the form of the arrival times is maintained. Below, we will show that propagation can easily adhere to this requirement and can be performed exactly and efficiently. However, performing an exact merging operation requires that all possible values of each of the random variable L_i in expression EQ11 be enumerated, which has an exponential run time in terms of the number of random variables L_i. This is computationally complex and also destroys the required form of arrival time. We therefore propose an alternate method for merging two arrival times, and prove that this method results in an arrival time whose CDF is an upper bound on the CDF of exact arrival times, while preserving the form of the arrival time expression. The method is simple and has linear run time with the number of random variables L_i. Using this approach, it is therefore possible to perform statistical timing analysis with linear run time in terms of circuit size, while guaranteeing a conservative analysis.

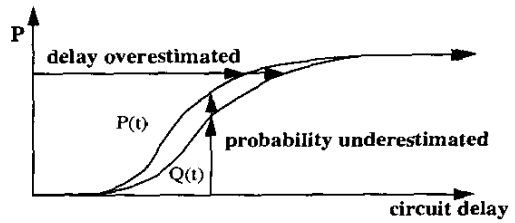

Figure 2. CDF $Q(t)$ is a conservative bound on CDF $P(t)$.

Below, we first define a statistical bound on the CDF of a random variable. We then discuss the methods for arrival time propagation and merging. Finally, in Section 4, we present a method whereby multiple arrival times can be propagated, improving the obtained bound at the cost of additional run time.

Statistical bounds

We define an upper bound on the CDF of an arrival time random variable as follows:

Definition 1. The arrival time CDF $Q(t)$ is an upper bound of the arrival time CDF $P(t)$ if and only if for all t, $Q(t) \le P(t)$.

Figure 2 shows two arrival time CDFs $P(t)$ and $Q(t)$, where $Q(t)$ is an upper bound on $P(t)$. Note that the upper bound $Q(t)$ is itself a valid CDF and that all confidence points are bounded by $Q(t)$ on $P(t)$. By using CDF $Q(t)$ instead of $P(t)$, we will overestimate the delay corresponding to a performance yield, resulting in a conservative analysis for late arrival times, as shown in Figure 2. Similarly, for a particular required delay, the probability that a die will meet this delay constraint will be underestimated.

We now introduce following useful lemma for arrival time CDFs:

Lemma A. If two random variables a and x have arbitrary CDFs $P(a)$ and $Q(x)$ and for any value of a random variable x is such that $a \le x$ then, the probability distribution of x is a statistical upper bound on the probability distribution of a.

Proof: Consider an arbitrary fixed value of t. We then separate cases $x \le t$ and $x > t$ and using the fact that according to the assumption $a \le x$, we can write:

$$\begin{aligned} P(a \le t) &= P(x \le t, a \le t) + P(x > t, a \le t) \\ &= P(x \le t) + P(x > t, a \le t) \end{aligned} \qquad \text{(EQ 13)}$$

From which it follows that

$$P_a(t) = P(x \le t) + P(x > t, a \le t) \quad \text{(EQ 14)}$$
$$= P_x(t) + P(x > t, a \le t) \ge P_x(t)$$

□

It should be noted that in Lemma A random variables a and x need not be statistically independent. We now show how arrival times can be computed using propagation and merging. While the propagation operation is exact, the merging operation results in an upper bound on the CDF of the exact arrival time.

Arrival time propagation

During the propagation operation, the delay of a gate is added to an arrival time. We perform this operation using the following procedure:

Procedure 1:

Given a gate delay $d = D_{nom} + \sum_i \alpha_i \cdot L_i + \Delta D_{random}$ and an arrival time $a_1 = A_{nom,1} + \sum_i \beta_{i,1} \cdot L_i + \Delta A_{random,1}$, at the input of the gate, we now compute the arrival time a_2 at the output of the gate as follows:

$$a_2 = A_{nom,2} + \sum_i \beta_{i,2} \cdot L_i + \Delta A_{random,2} \quad \text{(EQ 15)}$$

where
$$A_{nom,2} = D_{nom} + A_{nom,1}$$
$$\beta_{i,2} = \alpha_i + \beta_{i,1}$$
$$\Delta A_{random,2} = \Delta D_{random} + \Delta A_{random,1}$$

The derivation of the above procedure can be easily shown as follows: $a_2 = a_1 + d$. From this it follows that,

$a_2 = D_{nom} + \sum_i \alpha_i \cdot L_i + \Delta D_{random} + A_{nom,1} + \sum_i \beta_{i,1} \cdot L_i + \Delta A_{random,1}$

Simple rearranging of the terms results in $D_{nom} + A_{nom,1} + \sum_i (\alpha_i + \beta_{i,1}) \cdot L_i + \Delta D_{random} + \Delta A_{random,1}$, from which follows EQ15. Note that the computation of a_2 using EQ15 is exact and therefore correctly accounts for the spatial correlation of the arrival time a_1 and the gate delay d. Also, propagation using EQ15 is efficient as a simple summation of the coefficients of a_1 and d is performed. Since random variables ΔD_{random} and $\Delta A_{random,1}$ are independent, computation of $\Delta A_{random,2}$ is performed by simple numerical convolution.

Maximum operation

As mentioned earlier, computing an exact maximum of two arrival times a_1 and a_2 where each is expressed as a sum of components, requires enumeration of the random variables L_i, which is expensive. Also, the resulting arrival time would not be in the required form and spatial information would not be available for futher propagation and merging operations. We therefore propose a merging operation, which is efficient, and which generates an arrival time whose CDF is an upper bound on the exact arrival time. The proposed procedure is based on the following theorems.

Lemma B: For any given numbers a, b, x, y the inequality $max(a+b, x+y) \le max(a,x) + max(b,y)$ is valid

Proof: There exist only 4 mutually cases: (a) $a \le x, b \le y$; (b) $a \le x, b > y$; (c) $a > x, b \le y$ and (d) $a > x, b > y$.

In case (a) $max(a+b, x+y) = x+y$ and $max(a,x)+max(b,y)=x+y$ so inequality $max(a+b, x+y) \le max(a,x) + max(b,y)$ is valid.

In case (b) $max(a,x)+max(b,y) = x+b$ and according to the assumption both $a+b \le x+b$ and $x+y \le x+b$ are valid. So inequality $max(a+b, x+y) \le max(a,x) + max(b,y)$ is again valid.

Cases (c) and (d) are symmetrical to cases (b) and (a), proving the lemma. □

This lemma can be generalized to the following Theorem.

Theorem A: For any given numbers $a_1,a_2,..., a_n$ and $x_1,x_2,...,x_n$ the following inequality is valid.

$$max\left(\sum_{i=1}^{n} a_i, \sum_{i=1}^{n} x_i\right) \le \sum_{i=1}^{n} max(a_i, x_i). \quad \text{(EQ 16)}$$

Theorem A can be proven by induction using Lemma B.

Note that Theorem A holds for any numbers, regardless of their nature, including random variables. Applying Theorem A to the maximum of two arrival times, we can formulate the following procedure for the merge operation.

Procedure 2 :

Given arrival times, $a_1 = A_{nom,1} + \sum_i \beta_{i,1} \cdot L_i + \Delta A_{random,1}$ and $a_2 = A_{nom,2} + \sum_i \beta_{i,2} \cdot L_i + \Delta A_{random,2}$, we can compute an upper bound of merged arrival time a_3 as follows:

$$a_3 = A_{nom,3} + \sum_i \beta_{i,3} \cdot L_i + \Delta A_{random,3} \quad \text{(EQ 17)}$$

where
$$A_{nom,3} = max(A_{nom,1}, A_{nom,2})$$
$$\beta_{i,3} = max(\beta_{i,1}, \beta_{i,2})$$
or
$$\beta_{i,3} = min(\beta_{i,1}, \beta_{i,2})$$
$$\Delta A_{random,3} = max(\Delta A_{random,1}, \Delta A_{random,2})$$

Based on Theorem A, we can replace the maximum function $max(a_1, a_2)$ in procedure 2 with $max(A_{nom,1}, A_{nom,2})$, $max(\beta_{i,1} \cdot L_i, \beta_{i,2} \cdot L_i)$, and $max(\Delta A_{random,1}, \Delta A_{random,2})$. It is clear that $max(\beta_{i,1} \cdot L_i, \beta_{i,2} \cdot L_i) = max(\beta_{i,1}, \beta_{i,2}) \cdot L_i$, as shown in Procedure 2, for the positive values of the random variable L_i and $max(\beta_{i,1} \cdot L_i, \beta_{i,2} \cdot L_i) = min(\beta_{i,1}, \beta_{i,2}) \cdot L_i$, for the negative values of L_i. Also, since $\Delta A_{random,1}$ and $\Delta A_{random,2}$ are correlated only through path reconvergence, ignoring their correlation during their maximum computation will result in an upper bound [18], and hence the maximum of $\Delta A_{random,1}$ and $\Delta A_{random,2}$ can be efficiently computed numerically.

4 Multiple Arrival Time Propagation

While the maximum operation in Procedure 2 has the desired features that it is conservative and preserves the required form of arrival times, it nevertheless introduces error in the analysis. The degree to which error is introduced by Procedure 2 is dependent on the relative magnitude of the different components of a_1 and a_2. If, for instance, $\beta_{i,1} > \beta_{i,2}$ for all i, and also $A_{nom,1} > A_{nom,2}$ and the minimum value of $\Delta A_{random,1}$ with non-zero probability is greater than the maximum value of $\Delta A_{random,2}$ with non-zero probability (i.e. $\Delta A_{random,1} > \Delta A_{random,2}$ for all possible values), it is easy to show that the arrival time computed by Procedure 2 is exact. However, if some terms of arrival time a_1 are greater than a_2 and some terms of arrival time a_2 are greater than a_1, it is clear that a (conservative) error is introduced in the analysis.

To improve the analysis, we therefore extend the proposed appraoch by propagating multiple arrival times. In this case, only those arrival times are merged that result in a small error while those arrival times whose merger would result in a high error are propagated separately. If the correct arrival times are selected, it is clear that the analysis accuracy will improve. Given a set k of K arrival times incident at a node, we must select a subset m of M arrival times to propagate, while all other arrival times are merged with other arrival times. It is clear that the optimal set of arrival times to propagate depends on many factors, including the arrival times that will combine with the set m later in the circuit. Determining the optimal set is an intractable problem. We therefore propose the following heuristic to select the set of arrival times m given a set of incident arrival times k.

First, we compute for each pair of arrival times, m_i and m_j the maximum arrival time $a_{i,j}$ using Procedure 2. Then, we determine the mean of each arrival time $a_{i,j}$ which is computed by summing the means of each component. Finally, we select the arrival time $a_{i,j}$ with the minimum mean and replace the original two arrival times m_i and m_j with $a_{i,j}$ in the set m. This procedure reduces the size of the set m by one arrival time. The procedure is then repeated until the number of arrival times in m is reduced to a set of K arrival times, that can be propagated.

The above selective merging procedure effectively merges those arrival times incident on a node that result in an "early" arrival time that will have less impact on the overall delay of the circuit. These arrival times are therefore good candidates for merging, while arrival times whose merger would result in a late arrival time are propagated. The selective merging procedure is repeated at each node.

Finally, at the output node of the circuit, the set of K propagated arrival times must be merged to obtain the final arrival time of the circuit as a whole. Since the arrival times K do not need to be propagated further in the circuit, their particular form, in terms of a sum of independent random components, need not be preserved. Hence, we can convolve the components $A_{nom,i}$, $(\beta_{j,i} \cdot L_j)$, and $\Delta A_{random,i}$ into a single random variable before taking their maximum. This has the advantage that the error introduced by Procedure 2 is not incurred in the final merger of the arrival times at the output node. However, the arrival times are correlated, and to compute their exact maximum would require high computational complexity. We will therefore show that, due to the particular form of the arrival times, their correlation can be ignored and the computed maximum will bound the exact maximum. Hence, the maximum of the convolved arrival times can be efficiently computed using simple numerical techniques.

In [18], it was shown that the CDF of $max(x_1+y, x_2+z)$, where x_1, x_2, y, and z are independent random variables, is an upper bound on the CDF of $max(x+y, x+z)$, when x_1 and x_2 have an identical probability distribution as x. However, the form of our particular problem is more general in that we require the computation of $max(x+y, ax+z)$, where x, y, and z are independent random variables and a is a positive constant. We will now show that, similar to the previous case, the CDF of $\max(x_1+y, ax_2+z)$ is an upper bound on the CDF of $\max(x+y, ax+z)$. This means that ignoring the correlation between the two arrival times $(x+y)$ and $(ax+z)$ during the maximum operation will result in an upper bound of the CDF of the exact maximum. We prove the correctness of this bound with the following theorem.

Theorem 1: Let x, x_1, x_2, y, and z be positive, independent random variables with probability density functions $p(x)$, $p(x_1)$, $p(x_2)$ $q(y)$, $r(z)$, noting that x_1 and x_2 have the same probability density functions as random variable x. For any positive constant value $a>0$ the CDF of random variable $max(x+y, ax+z)$ is upper bounded by the CDF of random variable $max(x_1+y, ax_2+z)$.

Proof: The CDF of random variable $max(x+y, ax+z)$ is:

$$P(t) = \int_{max(x+y,\, a\cdot x+z)\le t} p(x)q(y)r(z)dxdydz \qquad \text{(EQ 18)}$$

The CDF of random variable $max(x_1+y, ax_2+z)$ is:

$$Q(t) = \int_{max(x_1+y,\, a\cdot x_2+z)\le t} p(x_1)p(x_2)q(y)r(z)dx_1dx_2dydz \qquad \text{(EQ 19)}$$

By transforming the integral over the 4 dimensional volume into an iterated integral we express $Q(t)$ as follows:

$$\int_0^\infty\int_0^\infty q(y)r(z)\left(\int_{x_1\le t-y,\, x_2\le \frac{t-z}{a}} p(x_1)p(x_2)dx_1dx_2\right)dydz \qquad \text{(EQ 20)}$$

We now rewrite $Q(t)$ as follows:

$$Q(t) = \int_0^\infty\int_0^\infty q(y)r(z)R(y,z)dydz\,, \text{ where} \qquad \text{(EQ 21)}$$

$$R(y,z) = \int_{x_1\le t-y,\, x_2\le \frac{t-z}{a}} p(x_1)p(x_2)dx_1dx_2 \qquad \text{(EQ 22)}$$

Multiplying equation EQ18 by the integral of probability density function $p(x)$ $\int_0^\infty p(v)dv = 1$ we express $P(t)$ as follows:

$$P(t) = \int_{max(x+y,\, a\cdot x+z)\le t} p(x)q(y)r(z)dxdydz \cdot \int_0^\infty p(v)dv \qquad \text{(EQ 23)}$$

We now rewrite $P(t)$ as follows, by rearranging the terms:

$$\int_{max(x+y,\, a\cdot x+z)\le t,\, v\le\infty} p(x)p(v)q(y)r(z)dxdvdydz \qquad \text{(EQ 24)}$$

We now convert this integral over the 4 dimensional volume into an iterated integral we obtain the following expression for $P(t)$:

$$\int_0^\infty \int_0^\infty q(y)r(z)\left(\int_{x \le t-y, x \le \frac{t-z}{a}, v \le \infty} p(x)p(v)dxdv \right) dydz \quad \text{(EQ 25)}$$

which can be rewritten in the form similar to EQ21:

$$P(t) = \int_0^\infty \int_0^\infty q(y)r(z)S(y,z)dydz \text{, where} \quad \text{(EQ 26)}$$

$$S(y,z) = \int_{x \le t-y, x \le \frac{t-z}{a}, v \le \infty} p(x)p(v)dxdv \quad \text{(EQ 27)}$$

This expression for $S(y,z)$ can be rewritten as follows:

$$S(y,z) = \int_{x \le min\left(t-y, \frac{t-z}{a}\right), v \le \infty} p(x)p(v)dxdv \quad \text{(EQ 28)}$$

The integrals expressing $R(y,z)$ and $S(y,z)$ in formulae EQ22 and EQ28 have the same integration functions $f(x_1,x_2)=p(x_1)p(x_2)$ and $f(x,v)=p(x)p(v)$ and differ only in the names of their variables. Moreover function $f(x,v)$ is symmetric with respect to its variables: $f(x,v) = f(v,x)$. Using this, we can prove that for any given values of y and z $R(y,z) \le S(y,z)$. We do this by considering two separate cases: $t-y \le (t-z)/a$ and $t-y > (t-z)/a$.

In the first case $min(t-y,(t-z)/a) = t-y$ and EQ28 becomes:

$$S(y,z) = \int_{x \le (t-y), v \le \infty} p(x)p(v)dxdv \quad \text{(EQ 29)}$$

Comparing EQ22 and EQ29 and renaming the integration variables x_1 and x_2 into x and v we can conclude that:

$$R(y,z) = \int_{x_1 \le t-y, x_2 \le \frac{t-z}{a}} p(x_1)p(x_2)dx_1dx_2 \le \int_{x_1 \le t-y, x_2 \le \infty} p(x_1)p(x_2)dx_1dx_2 = \int_{x \le t-y, v \le \infty} p(x)p(v)dxdv = S(y,z) \quad \text{(EQ 30)}$$

In the second case $min(t-y,(t-z)/a) = (t-z)/a$ and EQ28 becomes:

$$S(y,z) = \int_{x \le \frac{t-z}{a}, v \le \infty} p(x)p(v)dxdv \quad \text{(EQ 31)}$$

Comparing EQ22 and EQ31 and renaming integration variables x_1 and x_2 into v and x we can conclude that:

$$R(y,z) = \int_{x_1 \le t-y, x_2 \le \frac{t-z}{a}} p(x_1)p(x_2)dx_1dx_2 \le \int_{x_1 \le \infty, x_2 \le \frac{t-z}{a}} p(x_1)p(x_2)dx_1dx_2 = \int_{x \le \frac{t-z}{a}, v \le \infty} p(x)p(v)dxdv = S(y,z) \quad \text{(EQ 32)}$$

Thus for any y and z $R(y,z) \le S(y,z))$, from which using EQ21 and EQ26 we obtain $Q(t) \le P(t)$. Therefore, according to Definition 1, CDF $Q(t)$ of random variable $max(x_1+y, ax_2+z)$ is an upper bound of the CDF $P(t)$ of random variable $max(x+y, ax+z)$. □

5 Results

The statistical bound computation, as well as the proposed refinement method were implemented and tested on the synthesized version of ISCAS85 [20] benchmark circuits. Delay sensitivities were calculated for the standard cell library which used a 180 nm nominal device length. We used 3 levels of intra-die variation to model spatial correlation, as shown in Figure 3. Accordingly, each gate k was randomly allocated a location on a 4x4 grid, which determined the random variables associated with that gate along the hierarchy. Process variability information was used for different scenarios having a total standard deviation of 10%,14% and 15% from L_{nom}. The computed bounds were compared with Monte Carlo simulation and worst case analysis. Monte Carlo simulation was performed for 10,000 samples. The worst case analysis assumes the total variation to be inter-die variation and computes the 99% confidence point for the circuit delay CDF by setting ΔL_{inter} at its 99% point. For each gate length random variable, a Gaussian delay distribution truncated at the 3 sigma point, was used.

Table 1 shows the results for the bound computation and refinement using multiple arrival time propagation. A total standard deviation of 14% was divided among inter-die (5.7%), intra-die with spatial correlation (8.06%) and random intra-die variation (10%). For each circuit, the total number of nodes/edges (*column* 2) is shown. The 99% confidence points for worst case analysis (*column 3*), for single and multiple arrival times (*column* 4 & 5) and for Monte Carlo (*column* 6) is shown. The % error between the Monte Carlo results and our approach (*column* 7) was 2.98% on an average. Although we only report the 99% points in Table 1, the computed bounds are CDFs and allow the computation of other confidence points. Column 8 shows the runtime of our algorithm for 100 arrival times. For most circuits, the run time is very small with the maximum being 300 seconds.

Table 2 shows comparisons between 99% confidence points obtained by our algorithm using 100 arrival times and Monte Carlo simulation for two different variation scenarios. In (*Column* 2, 3 & 4) a total standard deviation of 10% was equally divided among inter-die, intra-die and random variations. The average error for all the circuits was 2.35%. The runtimes were small, not exceeding 300 seconds. In (*Column* 5, 6 & 7) a total standard deviation of 15% was again equally divided among the three components. Average error was 4.63% for all circuits and maximum runtime was 280 seconds.

Figure 3 shows the CDFs for the proposed upper bounds with

Table 1. Results for a total variation of 14%

Circuit		Results for the 99% confidence pt. (ns)					runtime (s)
name	nodes/ edges	worst case	1/20/50 arrival times	100a times	Monte Carlo	% error	
c17	13/20	0.30	0.28/0.28/0.28	0.28	0.28	0.00	4
c499	561/978	2.01	2.06/1.90/1.88	1.87	1.82	2.7	6
c432	214/379	2.26	2.23/2.09/2.07	2.07	2.03	1.9	6
c880	425/804	2.69	2.58/2.43/2.42	2.42	2.41	0.4	7
c1355	570/1071	2.68	2.66/2.51/2.49	2.48	2.41	2.9	12
c1908	466/858	3.82	3.70/3.59/3.58	3.58	3.41	4.9	25
c2670	1059/1731	2.63	2.44/2.35/2.35	2.35	2.34	0.4	25
c3540	991/1972	3.88	3.89/3.73/3.69	3.66	3.41	4.3	30
c5315	1806/3311	3.70	3.55/3.36/3.35	3.35	3.24	3.5	32
c6288	2503/4999	10.6	10.33/10.50/10.54	10.6	9.32	10.7	300
c7552	2202/3945	4.99	4.73/4.48/4.48	4.48	4.45	0.8	30

Table 2. Results for a total variation of 10% and 15%

Circuit name	10% variation			15% variation		
	99% pt. Our Approach/ Monte Carlo (ns)	% error	run-time (s)	99% pt. Our Approach/ Monte Carlo (ns)	% error	run-time (s)
c17	0.26/0.26	0.00	5	0.29/0.29	0.00	5
c499	1.81/1.76	2.80	10	2.02/1.91	5.75	12
c432	2.01/1.99	1.10	15	2.19/2.11	3.79	10
c880	2.39/2.38	0.42	15	2.55/2.51	1.59	14
c1355	2.40/2.35	2.10	25	2.66/2.52	5.55	20
c1908	3.47/3.35	3.50	28	3.82/3.58	6.70	25
c2670	2.31/2.30	0.43	30	2.45/2.44	0.40	25
c3540	3.51/3.35	4.70	30	3.90/3.57	9.24	32
c5315	3.25/3.18	2.21	35	3.61/3.40	6.17	37
c6288	9.87/9.11	8.30	300	10.85/9.85	10.1	280
c7552	4.41/4.39	0.45	35	4.71/4.63	1.72	34

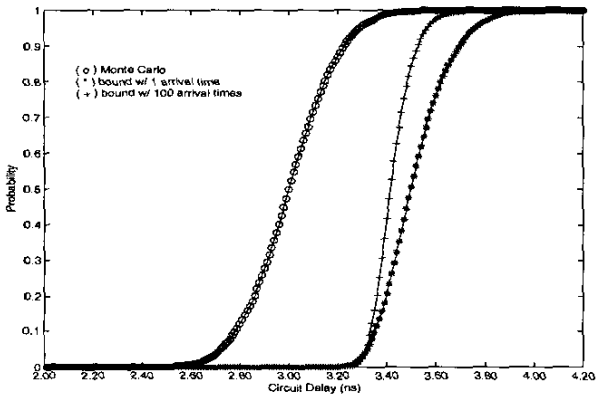

Figure 3. Comparison of CDF bounds and Monte-Carlo CDF

and without refinement as well as the CDF obtained through Monte Carlo simulation for the circuit c3540.

6 Conclusions

In this paper, we have proposed a new statistical timing analysis algorithm. The method has a linear run time and computes an upper bound on the distribution of the exact circuit delay. We first, proposed a model for inter- and intra-die process variations that accounts for spatial correlations. We then presented an efficient method for propagating arrival times in the circuit, which is linear in run time, and computes an upper bound on the distribution function of the exact circuit delay. We proved the correctness of the bound and showed how the bound is improved by propagating multiple arrival times at each node, using a heuristic method for selecting propagated arrival times. We tested the proposed methods on a number of synthesized benchmark circuits and demonstrated the accuracy and efficiency of the approach.

Acknowledgements

This research was supported by SRC, NSF, Intel and IBM.

References

[1] S. Nassif, "Delay Variability: Sources, Impacts and Trends," Proceedings of ISSCC, 2000.

[2] A. Kahng, Y. Pati, "Subwavelength optical lithography: challenges and impacts on physical design," Proceedings of ISPD, 1999.

[3] M. Orshansky, L. Milor, P. Chen, K. Keutzer, C. Hu, "Impact of systematic spatial intra-chip gate length variability on performance of high-speed digital circuits", ICCAD 2000, pp. 62 - 67.

[4] V.Mehrotra, S.L.Sam, D.Boning, A.Chandrakasan, R.Vallishayee, S.Nassif "A methodology for modelling the effects of systematic within-die interconnect and device variation on circuit performance. DAC 2000.

[5] Y. Deguchi, N. Ishiura, S. Yajima, "Probabilistic CTSS: analysis of timing error probability in asynchronous logic circuits," Proceedings IEEE/ACM Design Automation Conference, 1991.

[6] S. Devadas, H. F. Jyu, K. Keutzer, S. Malik, "Statistical timing analysis of combinational circuits ", ICCD 1992 pp. 38 -43

[7] H. F. Jyu, S. Malik, "Statistical timing optimization of combinational logic circuits" ICCD 1993. pp. 77 -80

[8] R.B. Brawhear, N. Menezes, C. Oh, L. Pillage, R. Mercer, "Predicting circuit performance using circuit-level statistical timing analysis" European Design and Test Conference, 1994.

[9] E.T.A.F. Jacobs, M.R.C.M Berkelaar, "Gate sizing using a statistical delay model," Proceedings IEEE/ACM Design Automation and Test Europe Conference, 2000, pp. 283-290.

[10] R.-B. Lin; M.-C. Wu, "A new statistical approach to timing analysis of VLSI circuits", Proc. Int. Conf. on VLSI Design, 1998

[11] S. Tongsima, C. Chantrapornchai, E.H.-M. Sha, N. L. Passos, "Optimizing circuits with confidence probability using probabilistic retiming," Proceedings IEEE ISCAS, 1998, pp. 270-273.

[12] L. Sheffer, "Explicit Computation of Performance as a Function of Process Variation", Int. Workshop on Timing Issues in the Specification and Synthesis of Digital Systems, TAU 2002.

[13] M. Berkelaar, "Statistical Delay Calculation, a Linear Time Method," Proceedings of TAU 97, Austin, TX, December 1997

[14] J.J Liou, K.T. Cheng, S. Kundu, A. Krstic, "Fast Statistical Timing Analysis By Probabilistic Even Propagation", DAC 2001

[15] M. Orshansky, K. Keutzer, "A general probabilistic framework for worst-case timing analysis", Proc. DAC 2002.

[16] A. Gattiker, S.Nassif, R.Dinakar, C.Long "Timing Yield Estimation from Static Timing Analysis," Proc., ISQED 2001

[17] X. Bai, C. Visweswariah, P. Strenski, D. Hathaway, "Uncertainty-aware circuit optimization," Proceedings ACM/IEEE Design Automation Conference, 2002.

[18] A. Agarwal, D. Blaauw, V. Zolotov, S. Vrudhula, "Computation and Refinement of Statistical Bounds on Circuit Delay," DAC 2003, pp. 348-353.

[19] A. Agarwal, D. Blaauw, S. Sundareswaran, V. Zolotov, M. Zhou, K. Gala, R. Panda, "Statistical Delay Computation Considering Spatial Correlations," ASP-DAC 2003, pp. 271-276.

[20] F. Brglez, H.Fujiwara, "A Neutral Netlist of 10 Combinatorial Benchmark Circuits", Proc. ISCAS, 1985, pp.695-698