

Statistical Static Timing Analysis With Conditional Linear MAX/MIN Approximation and Extended Canonical Timing Model

Lizheng Zhang, Weijen Chen, Yuhon Hu, and
Charlie Chung-ping Chen

Abstract—An efficient and accurate statistical static timing analysis (SSTA) algorithm is reported in this paper, which features 1) a conditional linear approximation method of the MAX/MIN timing operator, 2) an extended canonical representation of correlated timing variables, and 3) a variation pruning method that facilitates intelligent tradeoff between simulation time and accuracy of simulation result. A special design focus of the proposed algorithm is on the propagation of the statistical correlation among timing variables through nonlinear circuit elements. The proposed algorithm distinguishes itself from existing block-based SSTA algorithms in that it not only deals with correlations due to dependence on global variation factors but also correlations due to signal propagation path reconvergence. Tested with the International Symposium on Circuits and Systems (ISCAS) benchmark suites, the proposed algorithm has demonstrated very satisfactory performance in terms of both accuracy and running time. Compared with Monte-Carlo-based statistical timing simulation, the output probability distribution got from the proposed algorithm is within 1.5% estimation error while a 350 times speed-up is achieved over a circuit with 5355 gates.

Index Terms—Circuit performance analysis, extended canonical timing model, path reconvergence, process variation, statistical static timing analysis, very large scale integration (VLSI).

I. INTRODUCTION

The timing performance of deep submicrometer microarchitectures will be dominated by several factors. Integrated circuit (IC) manufacturing process parameter variations will cause device and circuit parameters to deviate from their designed value. Low supply voltage for low-power applications will reduce noise margin, causing increased timing delay variations. Due to dense integration and nonideal on-chip power dissipation, the rising temperature of a substrate may lead to hot spot, causing excessive timing variations. Classical worst case timing analysis produces timing predictions that are often too pessimistic and grossly conservative. On the other hand, statistical static timing analysis (SSTA) that characterizes timing delays as statistical random variables offers a better approach for more accurate and realistic timing prediction.

Existing SSTA methods can be categorized into two distinct approaches: path-based SSTA [1]–[4] and block-based SSTA [5]–[10]. Path-based SSTA seeks to estimate timing statistically on selected critical paths. However, the task of selecting a subset of paths whose time constraints are statistically critical has a worst case computation complexity that grows exponentially with respect to circuit size. Hence, path-based SSTA is not easily scalable to handle realistic circuits.

Manuscript received June 14, 2004; revised September 16, 2004 and February 8, 2005. This work was supported in part by Intel, TSMC, in part by UMC, in part by Faraday, in part by SpringSoft, in part by the National Science Foundation under Grant CCR-0093309 and Grant CCR-0204468, and the National Science Council of Taiwan, R.O.C., under Grant NSC 92-2218-E-002-030. This paper was recommended by Associate Editor D. Blaauw.

The authors are with the Department of Electrical and Computer Engineering, University of Wisconsin, Madison, WI 53706-1691 USA (e-mail: lizhengz@cae.wisc.edu; weijen@cae.wisc.edu; hu@engr.wisc.edu; chen@engr.wisc.edu).

Digital Object Identifier 10.1109/TCAD.2005.855979

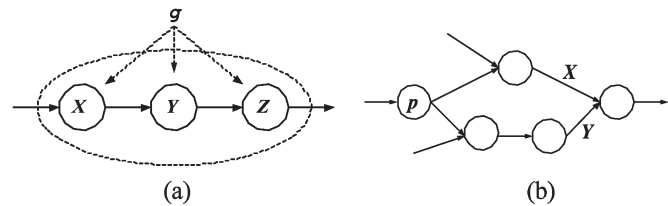


Fig. 1. Global correlations (left) and path correlation (right). (a) X , Y , and Z depend on g . (b) X and Y depend on p .

Block-based SSTA, on the other hand, champions the notion of progressive computation. Specifically, by treating every gate/wire as a timing block, SSTA is performed block by block in the forward direction in the circuit timing graph without looking back to the path history. As such, the computational complexity of block-based SSTA would grow linearly with respect to circuit size. However, to realize the full benefit of block-based SSTA, one must address a challenging issue that timing variables in a circuit could be correlated due to either global variations [6], [7], [10] or path reconvergence [5], [9]. As illustrated in the left-hand side of Fig. 1, global correlation refers to the statistical correlation among timing variables in the circuit due to global variations such as interdie or intradie spatial correlations, same gate type correlations, temperature or supply voltage fluctuations, etc. Path correlation, on the other hand, is caused by the phenomenon of path reconvergence, that is, timing variables in the circuit can share a common subset of gate/wire blocks along their path histories (Fig. 1).

The importance of path correlation comes from the fact that each gate/wire block in the circuit will have some local variations that are independent to the rest of the circuit. These local variations will propagate toward the circuit output and cause additional correlations due to the phenomenon of path reconvergence. Furthermore, these correlations caused by sharing local variations cannot be correctly captured by any algorithm that deals with global variations only. So for clarity, the term path correlation used here and after specifically refers to the correlation caused by the local variations of common path history.

Several solutions have been proposed to deal with either of these two types of correlations. In [6], [7], and [10], the dependence on global variations is explicitly represented using a canonical timing model. However, these approaches did not take into account path correlations. In [9], a method based on common block detection is introduced to deal with path correlations. However, this method does not address the issue of dependence on global variations. To the best of our knowledge, there is no existing method that has dealt with both types of correlations simultaneously. We present a novel block-based SSTA algorithm in this paper that is designed to consider both global correlations and path correlations.

- We develop a novel method to conditionally approximate the MAX/MIN operator by a linear mixing operator. Using the precomputed skewness, we are able to determine the linearity of the MAX/MIN operator analytically. Linear approximation is then applied only when MAX/MIN behaves linearly. When MAX/MIN is significantly nonlinear, the MAX/MIN evaluation is postponed with a form of max tuple.
- We extend the commonly used canonical timing model to be able to represent all possible correlations, including path correlations, between timing variables in the circuit. We further explore the sparse structure of the extended canonical representations of the timing variable and dynamically drop the nonsignificant terms so as to curtail the amount of storage and computation required for implementations.

Since $\min(X, Y) = -\max(-X, -Y)$, in the interests of brevity, in the rest of this paper, we will only discuss the MAX operator, with the understanding that the same results can be easily adapted to the MIN operator.

The rest of the paper is organized as follows. In Section II, previous block-based SSTA methods are reviewed briefly. Section III discusses the nonlinearity of the MAX operator and our conditional linear approximation method. Section IV describes the extended canonical timing (ECT) model and the proposed SSTA algorithm with the technique to reduce computation complexity. Section V presents a real implementation of our algorithm in C/C++ and the testing results with benchmark circuits. Section VI gives the conclusions.

II. BRIEF REVIEW OF CURRENT SSTA ALGORITHMS

For the purpose of timing analysis, timing blocks are used to represent the gate/wires in the circuit. Signals propagating through these blocks will add block delays into their arrival times. Block delays and arrival times are both called timing variables of the circuit. The history or path history of an arrival time is then defined as the set of block delays through which the signal ever passes.

A. Timing Variable Propagation

In statistical timing analysis, a timing variable is modeled as a random variable that is characterized by its distribution of probability density function (pdf) or, equivalently, cumulative distribution function (cdf). The goal of statistical timing analysis is to estimate the distribution of the arrival time in circuits given the distributions of each block delay in the circuit. To accrue the overall timing delay distribution, the timing delay random variables will be joined through two basic operators [5].

- ADD: When an input arrival time X propagates through a block delay Y , the output arrival time will be $Z = X + Y$.
- MAX: When two arrival times X and Y merge in a block, a new arrival time $Z = \max(X, Y)$ will be formulated before the block delay is added.

In the ADD operation, if both X and Y are Gaussian random variables, then $Z = X + Y$ will also be a Gaussian random variable whose mean and variance can be found as

$$\mu_Z = \mu_X + \mu_Y \quad (1)$$

$$\sigma_Z^2 = \sigma_X^2 + \sigma_Y^2 + 2\text{cov}(X, Y) \quad (2)$$

where $\sigma_{XY} = \text{cov}(X, Y) = E\{(X - \mu_X)(Y - \mu_Y)\}$ is the covariance between X and Y .

Denote $Z = \max(X, Y)$ to be the output of the MAX operator. Since MAX is generally a nonlinear operator, Z will not have a Gaussian distribution even if both X and Y are Gaussian. However, in this situation, the mean, variance, and skewness of the distribution of Z have been already derived analytically by Clark [11] as

$$\mu_Z = \mu_X Q + \mu_Y(1 - Q) + \theta P \quad (3)$$

$$\sigma_Z^2 = (\mu_X^2 + \sigma_X^2)Q + (\mu_Y^2 + \sigma_Y^2)(1 - Q) + (\mu_X + \mu_Y)\theta P - \mu_Z^2 \quad (4)$$

$$\begin{aligned} \kappa_Z^3 = \frac{1}{\sigma_Z^3} \left\{ (\mu_X^3 + 3\mu_X\sigma_X^2)Q + (\mu_Y^3 + 3\mu_Y\sigma_Y^2) \right. \\ \times (1 - Q) - \mu_Z(3\sigma_Z^2 + \mu_Z^2) \\ \left. + \frac{P}{\theta} \left((\mu_X^2 + \mu_X\mu_Y + \mu_Y^2)\theta^2 + 2\sigma_X^4 + \sigma_X^2\sigma_Y^2 \right. \right. \\ \left. \left. + 2\sigma_Y^4 - 2\sigma_{XY}(\sigma_X^2 + \sigma_Y^2) - \sigma_{XY}^2 \right) \right\} \quad (5) \end{aligned}$$

where $\theta = \sigma_{(X-Y)}$. P and Q are the pdf and cdf of a standard normal distribution evaluated at $\lambda = \mu_{(X-Y)}/\sigma_{(X-Y)}$, i.e.,

$$P(\lambda) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\lambda^2}{2}\right) \quad \text{and} \quad Q(\lambda) = \int_{-\infty}^{\lambda} P(x)dx. \quad (6)$$

The skewness of random variable Z is defined as

$$\kappa_Z = \frac{\sqrt[3]{E\{(Z - \mu_Z)^3\}}}{\sigma_Z}. \quad (7)$$

B. Linear Approximation of MAX Operator

Although $Z = \max(X, Y)$ does not have a Gaussian pdf even if both inputs X and Y are Gaussian distributed, in the interests of simplicity, it is still desirable by many SSTA researchers to find a Gaussian random variable that approximates Z in some way [6], [10]. In [6], the output of the MAX operator Z is approximated by a Gaussian random variable \hat{Z} , which is a linear combination of X , Y , and an additional independent Gaussian random variable Δ , i.e.,

$$Z = \text{MAX}(X, Y) \approx QX + (1 - Q)Y + \Delta = \hat{Z} \quad (8)$$

where Q is defined in (6) and is called tightness in [6]. The purpose of the additional random variable Δ is to ensure that the mean and the variance of \hat{Z} match those of Z as specified in Clark's formula (3) and (4).

In [11], it has also been shown that if W is a Gaussian random variable, then the cross-covariance between W and $Z = \text{MAX}(X, Y)$ can be found analytically as

$$\text{cov}(W, Z) = Q\text{cov}(W, X) + (1 - Q)\text{cov}(W, Y). \quad (9)$$

Substituting (8), it is easy to verify that

$$\text{cov}(W, \hat{Z}) = Q\text{cov}(W, X) + (1 - Q)\text{cov}(W, Y) = \text{cov}(W, Z).$$

Hence, a nice property of the approximator \hat{Z} shown in (8) is that the cross-covariance between Z and other timing variable W is preserved when Z is replaced by \hat{Z} .

While this approximation formula is simple, it does not work safely when the nonlinearity of the MAX operation is significant and the output of the MAX operator is significantly non-Gaussian. A simple example is illustrated in Fig. 2, where the left panel shows the two independent input Gaussian random variables and the right panel shows the cdfs of $\max(X, Y)$ from Monte Carlo simulation and linear approximation. It can be seen from Fig. 2(b) that the existing linear approximation will underestimate the distribution at high probability level. This behavior is risky since decisions made upon the estimated delay may result in excessive design failure.

C. Canonical Timing Model

Previously, a canonical timing model [6], [7], [10] has been proposed to address the delay correlations through shared global variations. In this model, the block delay is represented as a sum of three terms

$$n_i = \mu_i + \alpha_i R_i + \sum_{j=1} \beta_{i,j} G_j \quad (10)$$

where n_i ($i = 1, 2, \dots$) is the random variable corresponding to the i th block delay in the timing graph; μ_i is the expected value of n_i ; $R_i \sim N(0, 1)$ (called local variation) represents the localized statistical uncertainties of n_i ; $G_j \sim N(0, 1)$ represents the j th global

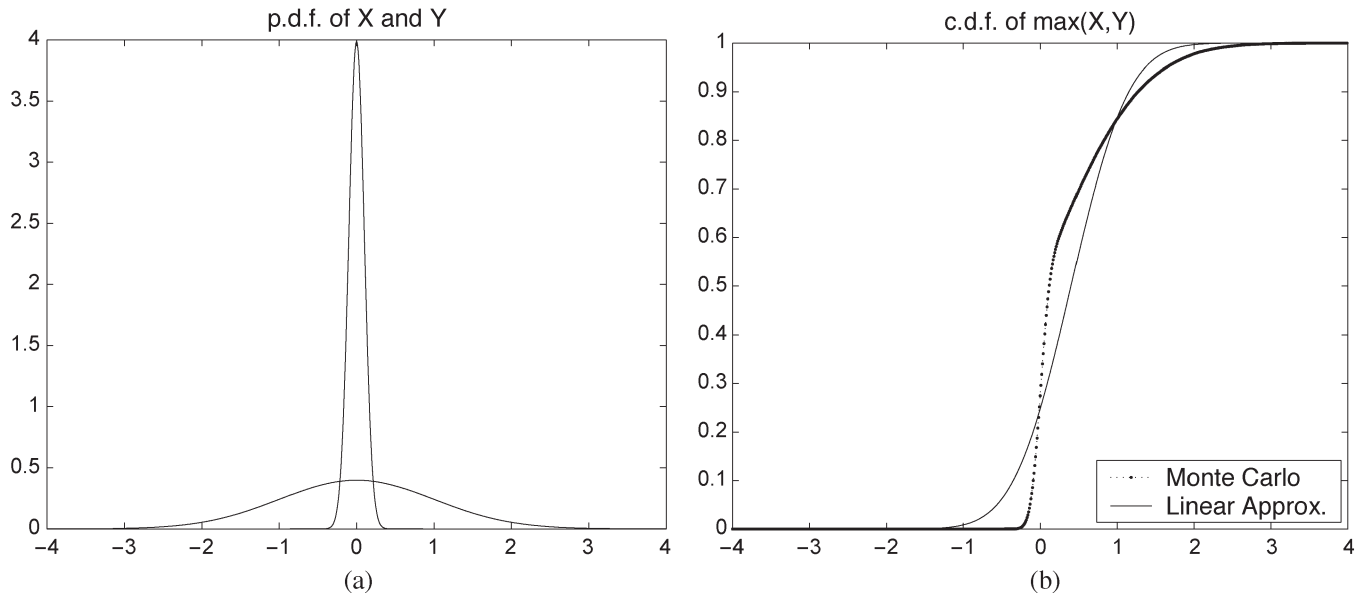


Fig. 2. Existing linear approximation underestimates MAX distribution at high probability level. (a) pdf of independent X and Y . (b) cdf of $\max(X, Y)$.

variation; R_i and $\{G_j (j = 1, 2, \dots)\}$ are additionally assumed to be mutually independent; and the weight parameters α_i (named local sensitivity) and $\beta_{i,j}$ (named global sensitivities) are deterministic constants, explicitly expressing the amount of dependence of n_i on each of the corresponding independent random variables.

With this canonical representation, the variance of a block delay n_i and its covariance with another block delay n_k can be evaluated as

$$\sigma_{n_i}^2 = E \{ (n_i - \mu_i)^2 \} = \alpha_i^2 + \sum_j \beta_{i,j}^2 \quad (11)$$

$$\text{cov}(n_i, n_k) = E \{ (n_i - \mu_i)(n_k - \mu_k) \} = \sum_j \beta_{i,j} \beta_{k,j}. \quad (12)$$

However, if arrival times are also expressed in this canonical model, the path correlation between them due to sharing local variations because of path reconvergence will incorrectly be ignored. For example, in Fig. 1(b), both arrival times X and Y include a common path history of block p . However, the local variation of block p , R_p , is no longer a part of the canonical representation of arrival times X and Y . Hence, the path correlation between X and Y due to R_p is incorrectly dropped.

III. NONLINEARITY OF MAX OPERATOR

For Gaussian inputs, the linearity of the MAX operator will be equivalent to the Gaussianity of the output. Using Monte Carlo simulation, the Gaussianity of the output can be evaluated with a method called QQ-Plot [12]. Specifically, if the output is Gaussian, then the simulated output of the MAX operator will show a straight line in its QQ-Plot against a standard Gaussian distribution. And if the MAX output is non-Gaussian, such QQ-Plot will deviate from linear. The more the non-Gaussianity of the MAX output, the worse the linearity of such QQ-Plot.

Since the linearity of the QQ-Plot can be quantitatively represented by the linear correlation coefficient of the QQ-Plot, the Gaussianity of the output of the MAX operator can be statistically and quantitatively measured. However, it will be very expensive if we run extensive Monte Carlo simulation during every step of MAX operation in timing analysis. So it is desirable to establish a more convenient criteria to determine the linearity of the MAX operator.

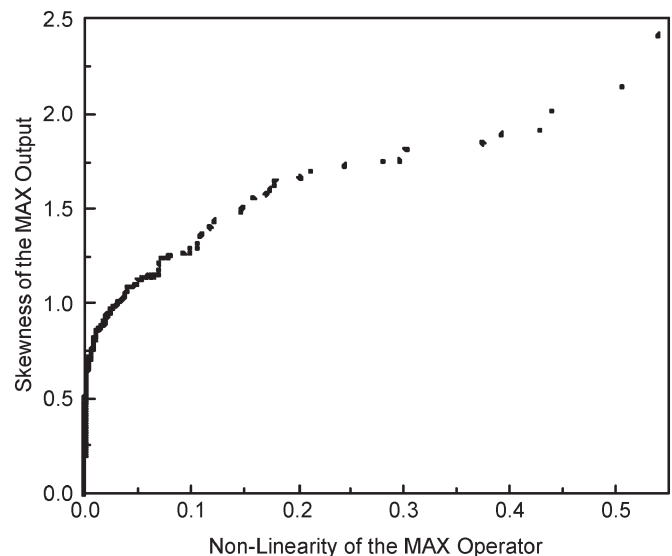


Fig. 3. Skewness of $Z = \max(X, Y)$ given X and Y are Gaussian versus the nonlinearity of MAX operator determined by Monte Carlo simulation.

It is well known that skewness is not a Gaussianity index for a general random variable since there are distributions that are symmetric but non-Gaussian. However, to measure the linearity of the MAX operator with Gaussian inputs, skewness of the MAX output will be a good choice. Fig. 3 shows the relationship between the nonlinearity of the MAX operator and the skewness of $Z = \max(X, Y)$ for Gaussian inputs X and Y . The scattering points in the figure represent 1000 random samples of the relative mean, relative variance, and the correlation of Gaussian random variables X and Y . The nonlinearity of the MAX operator for each set of randomly sampled mean, variance, and correlation is determined by QQ-Plot method with 10 000 Monte Carlo simulations. It is very clear in the figure that the skewness of the MAX output has a significant positive correlation with the nonlinearity of the MAX operator. Since skewness of the MAX output given Gaussian inputs can be analytically computed by equations developed by Clark [11], it is suitable to use skewness as an accurate and efficient measurement for the nonlinearity of the MAX operator.

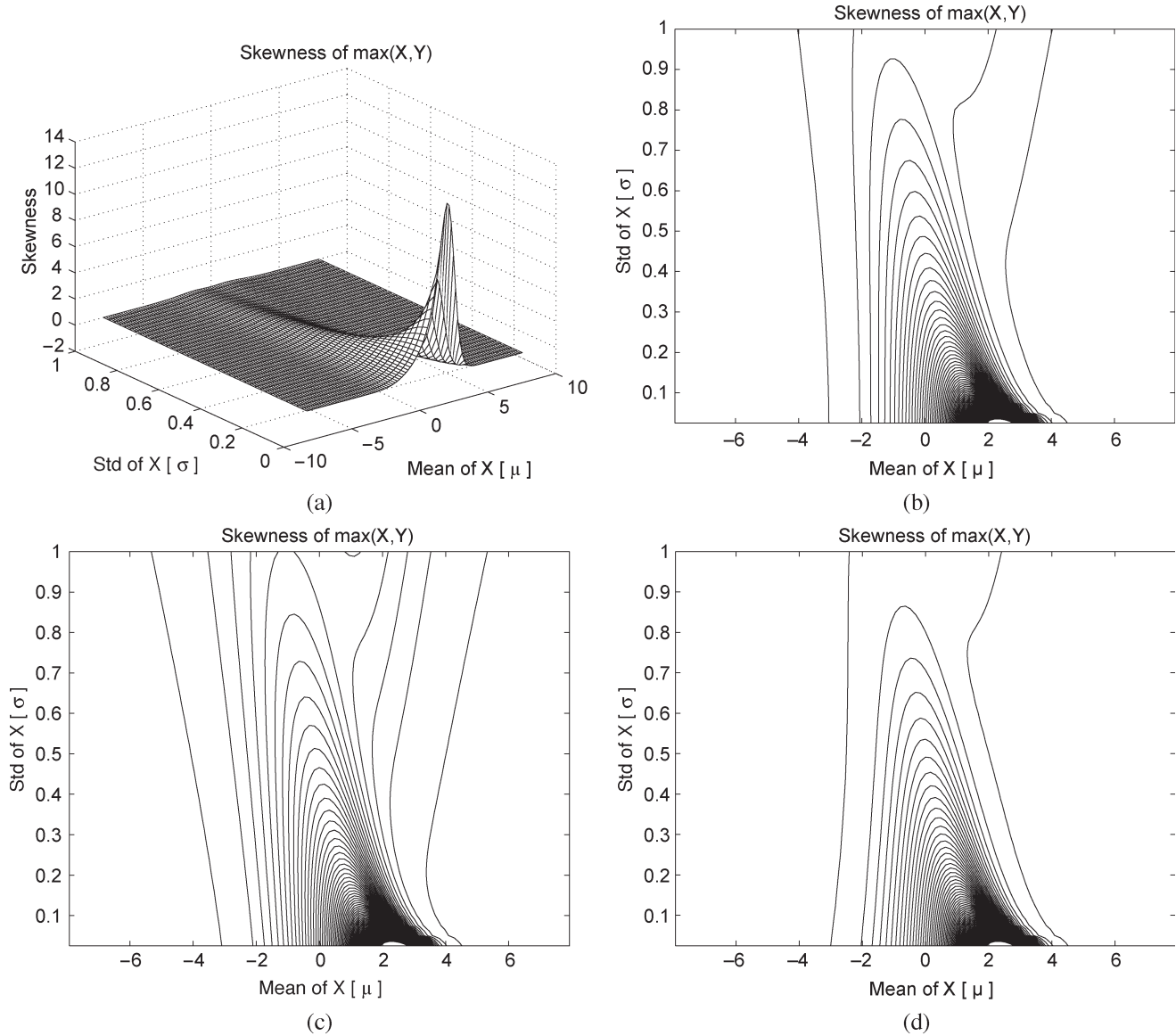


Fig. 4. Skewness of $\max(X, Y)$ when $Y \sim N(0, 1)$. (a) Three-dimensional (3-D) plot at $\rho = 0$. (b) Contour plot at $\rho = 0$. (c) Contour plot at $\rho = -0.5$. (d) Contour plot at $\rho = 0.5$.

A. Nonlinearity Condition of MAX Operator

It is clear that the linearity of the MAX operator is heavily dependent on its input parameters. Since we have a good measure of the linearity of the MAX operator, it is ready to study how the linearity changes when inputs vary.

Assuming the standard deviation $\sigma_Y \geq \sigma_X$ in $\max(X, Y)$, then no generality will be lost if the two variances are assumed to be $\sigma_X = \sigma \in [0, 1]$ and $\sigma_Y = 1$. This simplification is valid because of the scaling property of the MAX operator $\max(cX, cY) = c \max(X, Y)$ for any positive constant c . Aware of the invariance of the MAX operator in the constant shifting as $\max(X, Y) + c = \max(X + c, Y + c)$, both random variables of X and Y are shifted by the mean of Y so that the mean parameters will satisfy the range of $\mu_X = \mu$ and $\mu_Y = 0$. The last parameter that needed to specify the two input random variables involved in a MAX operation is their correlation coefficient ρ , which must be in the range of $-1 \leq \rho \leq 1$. With such parameter settings, the two Gaussian random variables X and Y are fully determined. And the skewness of $Z = \max(X, Y)$ are computed using equations developed by Clark [11] and are shown in Fig. 4.

From the figures, it is clear that in most of the cases the skewness is zero, which means $Z = \max(X, Y)$ is normally distributed and the MAX operator is linear. As a rule of thumb, the nonlinearity of the MAX operator is significant when the following nonlinear condition is satisfied.

Given X and Y are Gaussian, $\max(X, Y)$ will be significantly non-Gaussian if X and Y have very similar mean but very different variance or if X and Y have similar mean and variance but very negative correlation.

B. Conditional Linear MAX Approximation

Given two Gaussian random variables X and Y , $Z = \max(X, Y)$ could be significantly skewed if the nonlinear condition is satisfied. If the MAX operator is significantly nonlinear, significant error will occur if a linear operator is forced to approximate the MAX operator. But for the purpose of timing analysis, it is not necessary to explicitly compute the MAX output at every step.

1) *Max Tuple*: During timing analysis, the arrival time propagates from block to block with two elemental operations: ADD and MAX. If

during a propagation step of MAX $\max(X, Y)$ the output arrival time is not Gaussian, no actual computation will be done and the output will be simply recorded as a max tuple: $Mt\{X, Y\}$. With such max tuples, the arrival time propagation will have the following computations.

- ADD: Gate/wire delay D is added into a max tuple $Mt\{X, Y\}$ as

$$Mt\{X, Y\} + D = Mt\{X + D, Y + D\}.$$

- aMAX: Arrival time A is MAXed with a max tuple $Mt\{X, Y\}$ as

$$\max(A, Mt\{X, Y\}) = Mt\{A, X, Y\}.$$

- tMAX: Two max tuples are MAXed together

$$\max(Mt\{X, Y\}, Mt\{U, V\}) = Mt\{X, Y, U, V\}.$$

2) *Tuple Size*: To practically implement such tuple-based MAX evaluation, the number of arrival times in the max tuple, i.e., the tuple size, has to be maintained as small as possible. This is realized by the obvious combinational rule of max tuple as

$$\begin{aligned} Mt\{A, X, Y\} &= Mt\{\max(A, X), Y\} \\ &= Mt\{A, \max(X, Y)\} \\ &= Mt\{X, \max(A, Y)\} \end{aligned}$$

so if any two Gaussian random variables in the max tuple do not satisfy the nonlinear condition, then they can be replaced by a new Gaussian random variable by approximating the MAX with a linear operator and so that the size of the max tuple is reduced. This reduction process will be done iteratively to minimize the tuple size.

Such kind of tuple size reduction method is realized by associating each max tuple with a skewness matrix that stores the output skewness if pairs of random variables in the max tuple are actually MAXed out. And also a threshold of skewness κ_{th} is set beforehand to decide if the MAX result is Gaussian or non-Gaussian. Also, to prevent the explosion of the tuple size, a safeguard maximum allowed size for max tuple is also set, and if any of the tuple size exceeds the maximum size, the skewness threshold will be increased to tolerate more tuple size reduction.

Finally, in the primary output of the circuit, if the circuit delay is reported as max tuple, the output distribution can be easily evaluated by Monte Carlo simulation. For limited size of max tuple, such evaluation is efficient and accurate.

IV. ECT MODEL

The canonical timing model [6], [7], [10] is a powerful tool to represent the numerous timing variables for a given circuit. However, as pointed out in the previous section, in its original format, it can only handle timing correlations caused by global variations. In this work, we propose an ECT model that is capable of capturing all correlations between any pair of timing variables in the circuit, be it a block delay or an arrival time.

A. ECT Model

Assume that there are N gate/wire blocks and M global variations in the timing graph. If every block delay is modeled by the canonical

format shown in (10) and MAX is approximated by a linear combination operator, then every time variable, including all block delays and arrival times, will then have the ECT expression as

$$X = \mu_X + \sum_{i=1}^N \alpha_{X,i} R_i + \sum_{j=1}^M \beta_{X,j} G_j \quad (13)$$

where $R_i \sim N(0, 1)$ is the local and independent variation only related with block i , $G_j \sim N(0, 1)$ is the j th global variation, and $\alpha_{X,i}$ and $\beta_{X,j}$ are the corresponding sensitivity factors. To differ our approach from the existing canonical timing model, the word “extended” is used to indicate that the local variations are additionally included to the timing model. With such “extended” timing model, both global and path correlations can be handled elegantly. More specifically, global variations are represented by the set of global sensitivity terms $\{\beta_{X,j}\}$, and dependence on path history is represented by nonzero local sensitivity terms $\alpha_{X,k}$.

Equation (13) can be rewritten in a compacted vector format as

$$X \sim L(\mu_X, \alpha_X, \beta_X) = \mu_X + \alpha_X^* \mathbf{r} + \beta_X^* \mathbf{g} \quad (14)$$

where “*” means transpose and $\mathbf{r} \equiv [R_1, \dots, R_N]^* \sim N(\mathbf{0}, \mathbf{I})$ and $\mathbf{g} \equiv [G_1, \dots, G_M]^* \sim N(\mathbf{0}, \mathbf{I})$ are mutually independent local variation vector and global variation vector, respectively. $\mathbf{0}$ is a zero vector and \mathbf{I} is the unit matrix. $\alpha_X = [\alpha_{X,1}, \alpha_{X,2}, \dots, \alpha_{X,N}]^*$ and $\beta_X = [\beta_{X,1}, \beta_{X,2}, \dots, \beta_{X,M}]^*$ are deterministic local and global sensitivity vectors.

Chang and Sapatnekar [10] prove the correlation evaluation formula between timing variables represented by the canonical timing model of (10). We here prove a similar formula for correlation evaluation between time variables expressed with the ECT model as (13) or (14).

Theorem 1: Given timing variables $X \sim L(\mu_X, \alpha_X, \beta_X)$ and $Y \sim L(\mu_Y, \alpha_Y, \beta_Y)$, the correlation between them can be evaluated as

$$\text{cov}(X, Y) = \alpha_X^* \alpha_Y + \beta_X^* \beta_Y. \quad (15)$$

Proof: By definition

$$\begin{aligned} \text{cov}(X, Y) &= E\{(X - \mu_X)(Y - \mu_Y)\} \\ &= \text{cov}(\alpha_X^* \mathbf{r}, \alpha_Y^* \mathbf{r}) + \text{cov}(\alpha_X^* \mathbf{r}, \beta_Y^* \mathbf{g}) \\ &\quad + \text{cov}(\alpha_Y^* \mathbf{r}, \beta_X^* \mathbf{g}) + \text{cov}(\beta_X^* \mathbf{g}, \beta_Y^* \mathbf{g}) \\ &= E\{\alpha_X^* \mathbf{r} \mathbf{r}^* \alpha_Y\} + E\{\beta_X^* \mathbf{g} \mathbf{g}^* \beta_Y\} \\ &= \alpha_X^* \alpha_Y + \beta_X^* \beta_Y \end{aligned}$$

where the independence of \mathbf{r} and \mathbf{g} is applied. \blacksquare

To get the variance of a time variable, it is easy to prove the following corollary.

Corollary 1: Given timing variable $X \sim L(\mu_X, \alpha_X, \beta_X)$, its variance is

$$\sigma_X^2 = \alpha_X^* \alpha_X + \beta_X^* \beta_X \quad (16)$$

which is actually the special case when $X = Y$ of Theorem 1.

B. SSTA Algorithm

Before timing analysis, the delay sensitivities of each individual gate/wire are extracted from its Spice model and a gate/wire delay

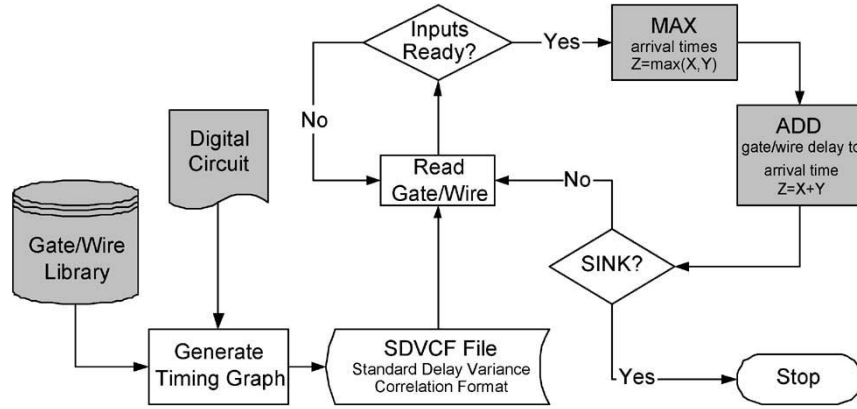


Fig. 5. Block-based SSTA algorithm.

library is then formed. This library, together with the circuit being analyzed, serves as the input of the SSTA algorithm. An SSTA algorithm will then calculate the distributions for all arrival times in the entire circuit by carrying out ADD and MAX operation at each gate/wire block. The overall data flow of the algorithm is summarized in Fig. 5, where the timing graph in the SSTA is represented by a file with standard delay variance correlation format (sdvcf), where both gate/wire delays and connections among gate/wires are specified.

Assuming $X \sim L(\mu_X, \alpha_X, \beta_X)$ and $Y \sim L(\mu_Y, \alpha_Y, \beta_Y)$, the output distribution of an ADD operation $Z = (X + Y) \sim L(\mu_Z, \alpha_Z, \beta_Z)$ can be easily computed as

$$\mu_Z = \mu_X + \mu_Y, \quad \alpha_Z = \alpha_X + \alpha_Y, \quad \beta_Z = \beta_X + \beta_Y. \quad (17)$$

According to the linear MAX approximation (8), the output distribution of the MAX operator $Z = \max(X, Y)$ will be

$$\begin{aligned} \mu_Z &= Q\mu_X + (1 - Q)\mu_Y + \theta P \\ \alpha_Z &= Q\alpha_X + (1 - Q)\alpha_Y \\ \beta_Z &= Q\beta_X + (1 - Q)\beta_Y. \end{aligned} \quad (18)$$

Clearly, the complexity of a single iteration of the SSTA algorithm comes from the sensitivity vector computation and the correlation evaluation involved in the MAX operation. Assuming there are totally M global variations and N gate/wire blocks in the circuit, the overall SSTA complexity will then be $O[(N + M)N]$.

C. Exploration of Sparsity

While working with benchmark circuits, we noticed that many components in the variation vectors have very small sensitivity values, indicating that their contributions to the overall correlation is insignificant. By setting these small coefficients to zero, the sensitivity vector will become a sparse vector that contains many zero components. Motivated by this observation, we apply a drop-and-lump method to exploit the sparsity of the sensitivity vector and to further decrease the average complexity of the SSTA algorithm.

For this purpose, a drop threshold is selected such that if $\alpha_{X,i}$ or $\beta_{X,j}$ is smaller than this threshold, it is deemed to have a small value and will be dropped from the sensitivity vector. However, dropping $\alpha_{X,i}$ or $\beta_{X,j}$ with a small magnitude directly is the same as applying truncation to the sensitivity vector. In subsequent computations, the quantization error may accumulate, causing nonnegligible error. This is a problem that cannot be overlooked for large circuits. Our solution

TABLE I
ERROR COMPARISON BETWEEN LUMPING AND DROPPING

Method	Simple Dropping	Lumping	MonteCarlo
τ_{97}	1343ps	1482ps	1431ps
Error	6.1%	3.4%	–

to this problem is to lump those dropped components into a single correction term

$$x_{\text{lump}} = \sqrt{\sum x_{\text{dropped components}}^2}. \quad (19)$$

Using circuit c499 as the example, the variation lumping method is compared with a simple dropping method at a drop level of 100%. From Table I, the advantage of using the lumping mechanism is clear: the estimation error for 97% delay quantile (τ_{97}) is improved from 6.1% of the simple dropping mechanism to 3.4% when the lumping mechanism is used.

Using this drop and lump method, the average number of nonzero terms in global sensitivity vector β will be M_C and the complexity of the proposed SSTA algorithm will be $O[(M_C + \Gamma)N]$, where the average number of nonzero terms in local sensitivity vector α is Γ . So what is really dropped in the local sensitivity vector α during computation is then the path correlation, and the length of the local sensitivity vector actually gives a good indication to the extent of path correlation in the circuit. So Γ is given a special name of path correlation length for a given drop threshold.

Theoretically speaking, if a block is not in any statistically critical path, its variation will be automatically dropped. On the other hand, if the block is in the critical path but is not statistically important, it will be dropped too. Furthermore, the importance of the variation will decrease after it propagates through a long path. In real circuits, usually only a few blocks in the circuit will survive the arrival time propagation so that $\Gamma \ll N$ and $M_C \ll N$. The computational complexity of the proposed method will be practically $O[(\Gamma + M_C)N] \approx O[N]$, and a significant reduction of complexity is achieved with the above drop and pool mechanism, although it is important to know that the actual complexity reduction is highly dependent on the topology of the circuit being analyzed.

V. SIMULATIONS AND DISCUSSIONS

Our SSTA algorithm, named as conditional linear MAX/MIN approximation with extended canonical timing model (CLECT), has already been implemented in C/C++ and tested by benchmark circuits. Before testing, however, all benchmark circuits are remapped into a

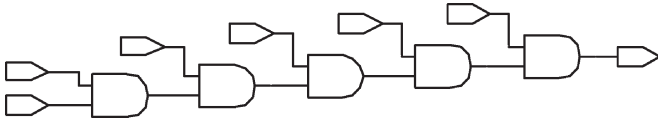


Fig. 6. Circuit whose timing variables are not Gaussian.

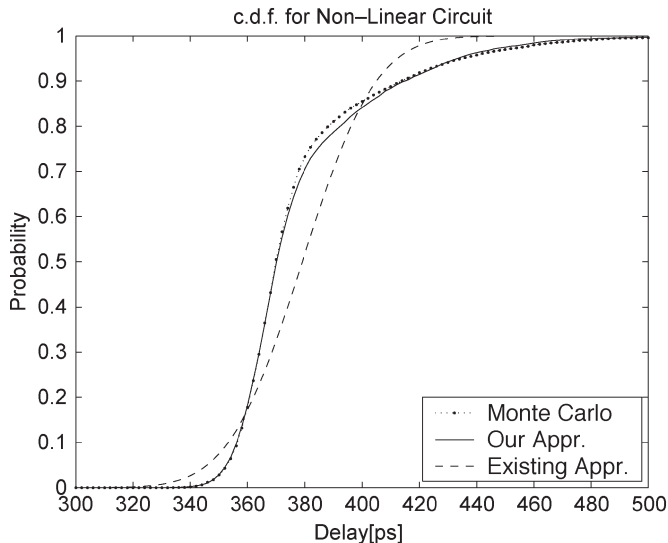


Fig. 7. Comparison of cdf for nonlinear circuit between 1) existing linear approximation and 2) conditional linear approximation with skewness threshold 0.5 and final tuple size of 3.

library that has gates of NOT, NAND2, NAND3, NOR2, NOR3, and XOR/XNOR. All library gates are implemented in 0.18- μm technology and their delays are characterized by Monte Carlo Spice simulation with Cadence tools assuming all variation sources follow Gaussian distribution.

For illustration purposes, only three parameter variations are considered global: channel length (L), supply voltage (V_{dd}), and temperature (T). All other variation sources, specified in the 0.18- μm technology file, are assumed to be localized in the considered gate only. Furthermore, we do not address the spatial dependency of the gate delays just for demonstration purposes. In real life, gate delay parameters are position dependent and our method is still applicable.

Extensive Monte Carlo simulations with 10000 repetitions are used as a “golden value” for each benchmark circuit. Each repetition is a process of static timing analysis by fixing global and block variations into a set of randomly sampled values. The global variations are sampled once for each repetition while block variation for each gate is sampled every time the gate delay is computed.

A. Accuracy Improvement With Max Tuple

One simple circuit is given in Fig. 6, where the overall delay and all internal MAX operators are significantly nonlinear as revealed by Monte Carlo simulation shown in Fig. 7. The skewness of the output distribution is $\kappa = 2.2$, which is significantly larger than zero. As shown in Fig. 7, it is very clear that for a circuit where MAX operators are significantly nonlinear, the existing linear approximation cannot correctly capture the cdf behavior. Especially, the existing method significantly underestimates the distribution at the high probability level so that it is risky to use such a distribution to predict circuit performance.

Our conditional linear MAX approximation method, on the other hand, matches the exact distribution much better than the existing method. Especially, at the high probability level, the computed distribution is almost exactly the same as the one got from Monte Carlo simulation. Such significant accuracy improvement over the existing method makes our method more suitable to predict performance for nonlinear circuits.

B. Accuracy Improvement by Including Path Correlation

Our SSTA method is also tested in the International Symposium on Circuits and Systems (ISCAS) benchmark suite. From Monte Carlo simulation, all ISCAS combinational circuits are Gaussian circuits whose MAX operators can be well approximated by linear operators. Both our conditional linear MAX approximation method and existing linear approximation will be good MAX approximation methods for these circuits. This nice property comes from the fact that for arrival times in these circuits, bigger mean usually means bigger variance and arrival times are usually positively correlated. So the nonlinear condition will not be satisfied for them.

Table II summarizes the error of the arrival time distribution parameters computed at the primary output for each testing circuit from three methods: 1) our method of CLECT; 2) NoPath, where the existing canonical timing model is used and no path correlation is considered; 3) NoCorr, where neither global correlation nor path correlation is considered. μ and σ are mean and standard variation of the distribution. $\tau_{97} = \mu + 2\sigma$ is the 97% delay quantile estimated assuming the output delay distribution is Gaussian.

From Table II, it is very clear that the method NoCorr fails to give reasonable variance estimation because no correlation is considered. This is a good example demonstrating the importance of correlations in SSTA. Table II also shows that the method NoPath has a significantly larger error in mean estimation than CLECT, although it shows similar accuracy in variance estimation. As a consequence, the method NoPath has a significantly larger error in 97% delay quantile estimation. This consistently larger error in all simulated circuits shows the importance to use the ECT model and consider the path correlations.

To further elaborate the accuracy improvement of CLECT over NoPath, Fig. 8 shows the pdf and cdf for circuit c6288 from three methods: Monte Carlo, CLECT, and NoPath. Apparently enough, CLECT shows excellent accuracy since it considers path correlation. And NoPath has a significant distribution shift because it uses canonical timing model and path correlation is dropped.

C. Performance and Path Correlation Length

It has been mentioned in Section IV-C that the path correlation length (Γ) is an interesting macroproperty of the simulated circuit and gives a good indication of the extent of path correlation existing in that circuit. For the above ISCAS circuits, the path correlation length (Γ) at a drop threshold of 1% is summarized in Table III, where the run time improvement over Monte Carlo simulation is also shown.

From Table III, we can first conclude that the correlation length Γ is much smaller than circuit size and basically independent on circuit size since it remains about 10–20 when the circuit size changes dramatically. This observation helps the conclusion we made before about the complexity reduction of our method by using the technique of flexible vector format.

Second, the only exceptionally high path correlation length among the tested circuits happens with the circuit c6288, which is known as a 16-bit array multiplier. Since there are a large amount of equal delay paths in the circuit, a large path correlation length is natural. Fewer local sensitivities can be dropped due to equal importance.

TABLE II
DISTRIBUTION ERROR RESPECTING TO MONTE CARLO RESULTS. 1) CLECT: OUR METHOD WITH EXTENDED CANONICAL MODEL. 2) NoPath: EXISTING CANONICAL MODEL WHERE NO PATH CORRELATION IS CONSIDERED. 3) NoCorr: NEITHER GLOBAL CORRELATION NOR PATH CORRELATION IS CONSIDERED

Circuit	Mean Error($\delta\mu$)			Variance Error($\delta\sigma$)			CPU Time(s)		
	CLECT	NoPath	NoCorr	CLECT	NoPath	NoCorr	CLECT	NoPath	NoCorr
c432	0.79%	4.64%	8.05%	0.50%	1.50%	89.8%	0.030	0.010	0.000
c499	1.04%	4.82%	6.99%	0.89%	0.34%	88.5%	0.030	0.010	0.000
c880	0.15%	1.22%	1.81%	0.53%	0.79%	93.8%	0.050	0.010	0.010
c1355	1.07%	5.81%	6.32%	0.28%	0.95%	95.0%	0.071	0.010	0.000
c1908	0.75%	2.93%	3.66%	0.27%	0.24%	91.8%	0.150	0.030	0.010
c2670	0.35%	3.09%	4.58%	0.00%	0.84%	94.3%	0.181	0.050	0.020
c3540	0.19%	3.75%	4.10%	0.66%	0.36%	95.3%	0.240	0.050	0.020
c5315	0.23%	3.12%	6.06%	0.11%	0.09%	92.8%	0.641	0.080	0.040
c6288	0.53%	8.17%	8.68%	0.65%	1.06%	98.8%	5.198	0.070	0.030
c7552	0.25%	3.27%	6.05%	1.46%	1.09%	92.5%	0.571	0.110	0.050

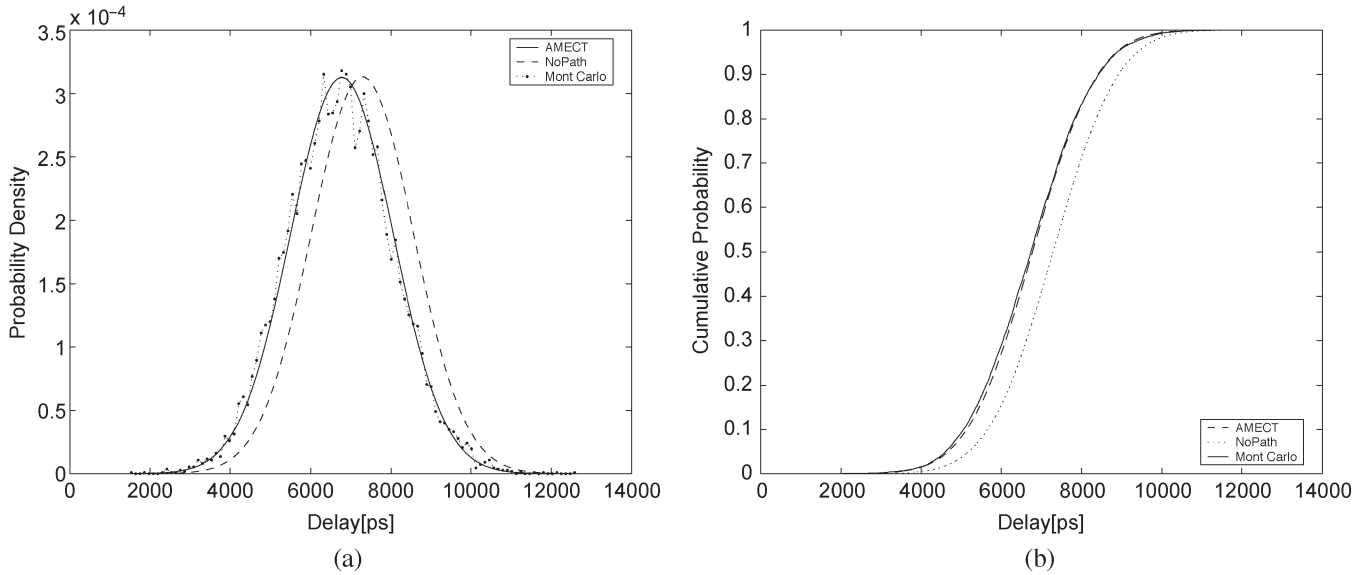


Fig. 8. pdf and cdf comparison for c6288 from three methods. (a) pdf from three methods. (b) cdf from three methods.

TABLE III
PATH CORRELATION LENGTH (Γ) AND CPU TIME IMPROVEMENT OVER MONTE CARLO SIMULATION

Name	c432	c499	c880	c1355	c1908
Gate Counts	280	373	641	717	1188
Γ	22.0	11.1	14.2	19.3	27.0
CPU Improve	217x	273x	297x	268x	239x
Name	c2670	c3540	c5315	c6288	c7552
Gate Counts	2004	2485	3865	2704	5355
Γ	15.4	21.2	14.4	80.9	16.0
CPU Improve	399x	350x	220x	22x	355x

To study the relationship between path correlation length and the accuracy of the SSTA method, an experiment is conducted for circuit c6288 and the results are shown in Fig. 9, where the errors in τ_{97} and path correlation length are both plotted against the drop threshold. It is clear that the path correlation length drops sharply when the drop threshold changes slightly from zero and maintains almost constant

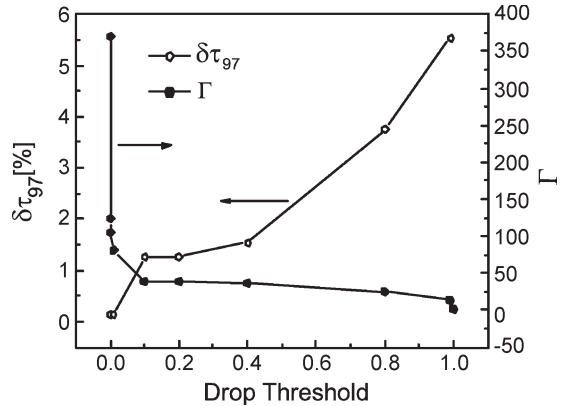


Fig. 9. Path correlation length (Γ) and error in 97% delay ($\delta\tau_{97}$) when drop threshold changes.

after that. But the error changes steadily when drop threshold changes. This phenomenon proves the efficiency of the drop mechanism introduced in this work since it means we can sacrifice very little accuracy

to gain very significant reduction in path correlation length and so as to save a significant amount of CPU time.

VI. CONCLUSION

This paper presents a novel method for block-based SSTA. We first disclose a new method to approximate the MAX operation with a linear operator assisted by skewness-based linearity evaluation. Second, we extend the commonly used canonical timing model into an "extended" version to represent the possible occurring path correlation. With these theoretical progress, we are able to evaluate and propagate both global and path correlation in the circuit timing graph. We also design a novel algorithm CLECT that treats both global and path correlation simultaneously and systematically. This algorithm, with the help of a drop-and-lump method, achieves high accuracy and high performance at the same time as tested by ISCAS circuits and compared with Monte Carlo "golden values."

ACKNOWLEDGMENT

The authors thank Prof. B. D. Van Veen for great discussions.

REFERENCES

- [1] J.-J. Liou, A. Krstic, L.-C. Wang, and K.-T. Cheng, "False-path-aware statistical timing analysis and efficient path selection for delay testing and timing validation," in *Proc. 39th Design Automation Conf.*, New Orleans, LA, Jun. 2002, pp. 566–569.
- [2] M. Orshansky and A. Bandyopadhyay, "Fast statistical timing analysis handling arbitrary delay correlations," in *Proc. 41st Design Automation Conf.*, 2004, pp. 337–342.
- [3] M. Orshansky and K. Keutzer, "A general probabilistic framework for worst case timing analysis," in *Proc. 39th Design Automation Conf.*, New Orleans, LA, Jun. 2002, pp. 556–561.
- [4] A. Agarwal, D. Blaauw, V. Zolotov, S. Sundareswaran, M. Zhao, K. Gala, and R. Panda, "Statistical delay computation considering spatial correlations," in *Proc. ASP-DAC*, Kitakyushu, Japan, Jan. 2003, pp. 271–276.
- [5] A. Agarwal, V. Zolotov, and D. Blaauw, "Statistical timing analysis using bounds and selective enumeration," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 22, no. 9, pp. 1243–1260, Sep. 2003.
- [6] C. Visweswariah, K. Ravindran, K. Kalafala, S. G. Walker, and S. Narayan, "First-order incremental block-based statistical timing analysis," in *Proc. 41st Design Automation Conf.*, 2004, pp. 331–336.
- [7] A. Agarwal, D. Blaauw, and V. Zolotov, "Statistical timing analysis for intra-die process variations with spatial correlations," in *Proc. ICCAD*, San Jose, CA, Nov. 2003, pp. 900–907.
- [8] S. Bhardwaj, S. B. Vrudhula, and D. Blaauw, " τ AU: Timing analysis under uncertainty," in *Proc. ICCAD*, San Jose, CA, Nov. 2003, pp. 615–620.
- [9] A. Devgan and C. Kashyap, "Block-based static timing analysis with uncertainty," in *Proc. ICCAD*, San Jose, CA, Nov. 2003, pp. 607–614.
- [10] H. Chang and S. S. Sapatnekar, "Statistical timing analysis considering spatial correlations using a single pert-like traversal," in *Proc. ICCAD*, San Jose, CA, Nov. 2003, pp. 621–625.
- [11] C. Clark, "The greatest of a finite set of random variables," *Oper. Res.*, vol. 9, no. 2, pp. 145–162, Mar. 1961.
- [12] P. Lewis and E. Orav, *Simulation Methodology for Statisticians, Operations Analysts, and Engineers*. London, U.K.: Chapman & Hall, 1988.