Statistical Timing Analysis with AMECT: Asymptotic MAX/MIN Approximation and Extended Canonical Timing Model

***Abstract*— State of the art statistical timing analysis (STA) tools often yield less accurate results when timing variables become correlated due to global source of variations and path reconvergence. To the best of our knowledge, no good solution is available dealing *both* types of correlations *simultaneously*.**

In this paper, we present a novel statistic timing algorithm, $AMECT$ (Asymptotic MAX/MIN approximation & Extended Canonical Timing model), that produces accurate timing estimation by solving *both* correlation problems simultaneously. Specifically, AMECT uses a linear mixing operator to approximate the nonlinear MAX/MIN operator by moment matching and develops an extended canonical timing model to evaluate and decompose correlations between arbitrary timing variables. Finally, AMECT is implemented by an intelligent pruning method to enable trade-off runtime with accuracy.

Tested with ISCAS benchmark suites, AMECT shows both high accuracy and high performance compared with Monte Carlo simulation results: with distribution estimation error $< 1.5\%$ while with around 350X speed up on a circuit with 5355 gates.

I. Introduction

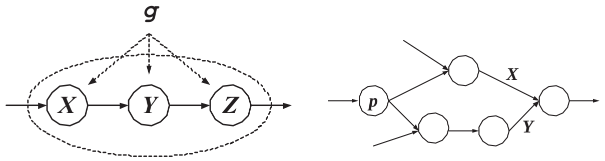
It is well-known that the timing performance of future generations of deep-submicron micro-architecture will be dominated by several factors. IC manufacturing process parameter variations will cause device and circuit parameters to deviate from their designed value. Low supply voltage for low-power applications will reduce noise margin, causing increased timing delay variations. Due to dense integration and non-ideal on-chip power dissipation, rising temperature of 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 timing analysis (STA) that characterizes timing delays as statistical random variables offers a better approach for more accurate and realistic timing prediction.

In literatures, there are two distinct approach for STA: **path based STA** and **block based STA**. The fundamental challenge of the path based STA [1]–[4] is its requirement to select a proper subset of paths whose time constraints are statistically critical. This task has a computation complexity that grows exponentially with respect to the circuit size, and hence can not be easily scaled to handle realistic circuits.

This potential difficulty has motivated the development of block base STA [5]–[10] that champions the notion of *progressive computation*. Specifically, statistical timing analysis is performed block by block in the forward direction in the circuit timing graph without looking back to the path history. As such, the computation complexity of block based STA will grow linearly with respect to the circuit size. To even further speed up the computation, *Gaussian assumption* has been widely adopted([6], [9], [10]) with small accuracy penalty, and all internal timing random variables in a circuits are forced to follow the Gaussian distribution.

However, to realize the full benefit of block based STA, one must solve a difficult problem that timing variables in a circuit could be correlated due to either *global variation* ([6], [7], [10]) or *path reconvergence*([5], [9]). As illustrated in the left hand side of Figure 1, *global correlation* refers to the statistical correlation among timing variables in the circuit due to *global variations* such as inter- or intra-die spatial correlations, same gate type correlations, temperature or supply voltage fluctuations, etc. *Path correlation*, illustrated in the right hand side of Figure 1, refers to the correlation resulting from the phenomenon of *path reconvergence*, that is, timing variables may share a common subset of gate or interconnect along their path histories.



(a) X, Y and Z depend on g (b) X and Y depend on p

Fig. 1. Global Correlations (left) and Path Correlation(right)

Several preliminary solutions have been proposed to deal with these correlations. In [6], [7], [10] the dependence on global variations is explicitly represented using a *canonical timing model*. In [6], an intuitively defined parameters, *tightness* is proposed to help the propagation of global correlations. However, none of these approaches has taken into account the path correlations. and the intuitively defined tightness parameter in [6] may find difficulty to accurately propagate the correlation information. In [9], a method based on common node detection is introduced to deal with the path correlations. However, this method does not address the issue of dependence on global variations.

In this paper, we present a systematic STA solution, named **AMECT**, that takes into account correlations caused by *both* global variations and path reconvergence. Specifically,

- We *extend* the commonly used canonical timing model to represent all timing variables in the circuit as a weighted linear combination of a set of independent random variables. A *variation vector*, consisting of all these weights, is then used to explicitly represent *both* global and path correlation information.

- We develop a novel method to decompose the correlations between timing variables and approximate the output of a nonlinear MAX/MIN operator by a linear mixing operator. As such, the variation vector can then be easily updated to retain the correlation information *after* the MAX/MIN operator.
- We further explore the sparse structure of the variation vector and develop a *flexible vector format* so that the non-significant entries of the variation vector are dynamically dropped during computation. According to simulations on ISCAS circuits, this technique significantly curtails the amount of storage and computation required for **AMECT** implementation.

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 following: In section II, previous block based STA methods are reviewed briefly; Sections III states and proves the MAX linearization theorem; Section IV describes the vectorized timing format and a theorem used for correlation decompose; Section V is the detailed algorithm and technique to reduce computation complexity. Section VI presents a real implementation of **AMECT** in C/C++ and the testing result with ISCAS85 benchmark suites; Section VII gives the conclusions.

II. A Brief Review of Current STA Algorithms

In timing analysis field, the circuit is modeled as a *timing graph*, which is a directed acyclic graph(DAG) where each delay source, including both logic gates and interconnects, is represented as a *node*. Each node connects to other nodes through some input and output *edges*. Nodes and edges are called *delay elements*. Each node is assigned with a *node delay* representing the delay incurred in the corresponding logic gates or interconnect segments. The *edge delay*, a short term of signal arrival time at the edge, represents the cumulative timing delays upto and including the node that feeds into the edge. The *history* or *path history* of the edge delay is then defined as the set of node delays through which the signal arrives at this edge ever passes.

Different from classical timing analysis, the statistical timing analysis models delay elements as *random variables*, which are characterized by its *probability density function(p.d.f.)* or *cumulative distribution function(c.d.f.)*. The purpose of statistical timing analysis is then to estimate the edge delay distribution at the primary output of the circuits knowing input edge delay distributions and all internal node delay distributions. This is accomplished through two operators [5]:

- ADD: When an input edge delay X propagates through a node delay Y, the output edge delay will be $Z = X+Y$
- MAX: When two edges delays X and Y merge in a node, a new edge delay $Z = max(X, Y)$ will be computed before the node delay is added.

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

$$\begin{aligned} \mu_Z &= \mu_X + \mu_Y & (1) \\ \sigma_Z^2 &= \sigma_X^2 + \sigma_Y^2 + 2cov(X, Y) & (2) \end{aligned}$$

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

Due to the nonlinearity, the output delay element of the MAX operator, $Z = max(X, Y)$, will not have Gaussian distribution even when both inputs are Gaussian distributed. For this case, Clark [11] in 1961 derived the first and second moments of the distribution of $max(X, Y)$: if X and Y are Gaussian and statistically independent,

$$\begin{aligned} \mu_Z &= \mu_X \cdot Q + \mu_Y(1 - Q) + \theta P & (3) \\ \sigma_Z^2 &= (\mu_X^2 + \sigma_X^2)Q + (\mu_Y^2 + \sigma_Y^2)(1 - Q) \\ &\quad + (\mu_X + \mu_Y)\theta P - \mu_Z^2 & (4) \end{aligned}$$

where $\theta^2 = \sigma_X^2 + \sigma_Y^2$. P and Q are $p.d.f.$ and $c.d.f.$ of standard Gaussian distribution at $\lambda = (\mu_X - \mu_Y)/\theta$:

$$P(\lambda) = \frac{1}{\sqrt{2\pi}} exp(-\frac{\lambda^2}{2}) \qquad Q(\lambda) = \int_{-\infty}^{\lambda} P(x)dx$$

When X and Y are correlated, similar, yet more complicated expressions for these moments have also been derived in [11].

An intuitive solution to the non-Gaussian problem of MAX operator is to use a Gaussian *p.d.f.* to approximate the MAX output distribution such that the first two moments of the Gaussian *p.d.f.* match those derived by Clark. This approach has been adopted in [6], [10]. Nonetheless, they fail to address the issue of path correlations among delay elements.

A. Canonical Timing Model

[6], [7], [10] proposed a *canonical delay model* to address the node delay correlations through sharing global variations. In particular, they model each of the node delay as a summation of three terms:

$$n_i = \mu_i + \alpha_i R_i + \sum_{j=1} \beta_{i,j} G_j \tag{5}$$

where $n_i (i = 1, 2, ...)$ are random variables corresponding to the the i^{th} node delay in the timing graph; μ_i is the expected value of n_i; R_i, (named *node variation*), is a zero-mean, unity variance Gaussian random variable representing the localized statistical uncertainties of n_i; G_j represents the j^{th} *global variation*, and is also modeled as a zero-mean, unity variance Gaussian random variable; $\{R_i\}$ and $\{G_j\}$ are additionally assumed to be mutually independent; the weight parameters α_i (named *node 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 correlation (covariance) between any two node delays, n_i and n_k, can be easily evaluated.

$$cov(n_i, n_k) = E\{(n_i - \mu_i)(n_k - \mu_k)\} = \sum_j \beta_{i,j}\beta_{k,j} \quad (6)$$

Note that random variables $\{R_i, R_k, G_j(j = 1, 2, ...)\}$ are mutually independent.

B. Existing Method for Handling Correlations

Delay elements in a timing graph, including node delays and edge delays, may become correlated due to sharing global variations and/or common path histories. Multiple methods handling one of these two types of correlations have been proposed to get more accurate STA estimation.

In [6], [7], [10], the canonical timing model of Equation (5) is directly applied into the edge delays in a timing graph. This direct usage implicitly assumes that edge delay only depends on global variations and no path correlation occurs in the timing graph. This method will work well apparently only when global variation dominates the correlations in the timing graph but will have severe problem where path correlation is important.

The authors in [6] propose the use of *tightness* to retain global correlation information through the nonlinear MAX operation. The global sensitivities of the output edge delay from a MAX operation is treated as a tightness-based supposition of the global sensitivities of the input edges delays. This method is valuable since it hints to use linear supposition as the replacement of nonlinear MAX operation. But using the intuitively defined tightness as the supposition coefficient is not a suitable choice as revealed in section III.

In [9], a common node detection procedure is introduce to deal with the path correlation. This method assumes that if two edge delays, X and Y, ever pass a common node whose output edge delay is W, then $X = X' + W$ and $Y = Y' + W$. Operation $max(X, Y)$ is then done as $W + max(X' + Y')$. This is not a good approximation since X and Y usually don't have such a strong dependence on W. A counter example is illustrated in Figure 2 where both X and Y are theoretically dependent on W. But practically speaking, X will be independent on W if $U >> W$ and similarly Y will be independent on W if $V >> W$.

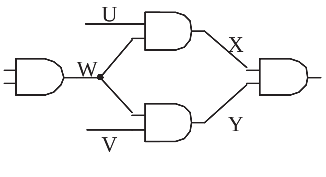


Fig. 2. Example to Fail Common Node Approach

To the best of our knowledge, existing STA methods have yet to offer a solution to deal with the correlation problem caused by *both* global parameter variations and path reconvergence.

III. Linear Mixing Approximation of MAX

In some of the current STA tools, the output of a MAX operator is approximated by a Gaussian distribution with its first two moments matching those derived by Clark (Equations (3 and 4)). However, the resulting Gaussian distribution will lose most of the correlation information between the input edge delays.

In this paper, instead of generating the Gaussian distribution at the MAX output directly, we propose to model the nonlinear MAX operator with a weighted linear mixing operator. Clearly, if inputs to the linear mixing operator are Gaussian distributed, so well the output after the linear mixing operation.

More specifically, we choose the weights of the linear mixing so that the first two moments of the resulting output Gaussian distribution match those derived by Clark (cf. equation (3 and 4)). While our method produces the same approximated distribution at the output of a MAX operator, the weighted linear mixing formulation makes it possible to retain the correlation information using an *extended canonical timing model* discussed in the next section. By preserving the correlation information after the nonlinear MAX operators, the accuracy of the STA can be significantly improved.

Theorem 1 (Max Linearization): *Let X, Y and Z be Gaussian random variables and that $cov(X, Y) = 0$. If the first two moments of Z match those of the random variable $max(X, Y)$, then there must exist a constant, ρ, $0 < \rho < 1$, called the contribution factor, such that:*

$$Z = \rho \cdot X + (1 - \rho) \cdot Y + \zeta \quad (7)$$

where ζ is an arbitrary constant.

Proof: Since random variable Z is a Gaussian approximation of $max(X, Y)$, it will then be fully determined by its first two moments given in Clark Equations If another Gaussian random variable $Z' = \rho \cdot X + (1 - \rho) \cdot Y + \zeta$ satisfies the following two moment matching equations:

$$\begin{aligned} \mu_Z &= \mu_{Z'} = \rho \cdot \mu_X + (1 - \rho)\mu_Y + \zeta && (8) \\ \sigma_Z^2 &= \sigma_{Z'}^2 = \rho^2\sigma_X^2 + (1 - \rho)^2\sigma_Y^2 && (9) \end{aligned}$$

then Z and Z' must be identical.

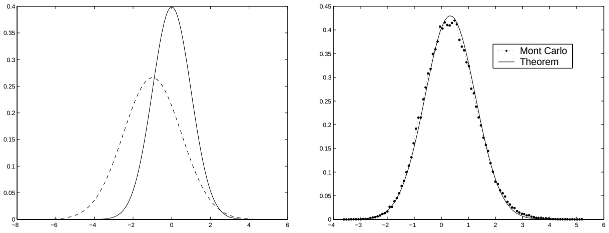
Solution to the mean matching Equation (8) will always exist since it is only a linear equation. So the proof becomes to guarantee real solutions for the quadratic variance matching Equation (9). This is equivalent to confirm:

$$\sigma_Z^2 \geq \frac{\sigma_X^2\sigma_Y^2}{\sigma_X^2 + \sigma_Y^2} \quad (10)$$

Using Clark's equations (3 and 4), the above inequality (10) is proved in Appendix I and solutions to Equation (9) will be:

$$\rho_\pm = \frac{\sigma_Y^2}{\sigma_X^2 + \sigma_Y^2} \pm \sqrt{\left(\frac{\sigma_X^2}{\sigma_X^2 + \sigma_Y^2}\right)^2 + \frac{\sigma_Z^2 - \sigma_X^2}{\sigma_X^2 + \sigma_Y^2}} \quad (11)$$

Also proved in Appendix II, if $\sigma_Y^2 \geq \sigma_X^2$, $0 < \rho_- < 1$. If $\sigma_Y^2 < \sigma_X^2$, $0 < \rho_+ < 1$. So by switching the contribution factor ρ between ρ_- and ρ_+ according to the relative magnitude of σ_X^2 and σ_Y^2, $0 < \rho < 1$ can always be guaranteed. ■



(a) $p.d.f.$ of X and Y

(b) $p.d.f.$ of $max(X, Y)$

Fig. 3. Mont Carlo and Theorem Result Comparison

In Figure 3(a), the *p.d.f.*s of two independent Gaussian random variables, X and Y are shown in the left panel. The *p.d.f.* from Monte Carlo method for a MAX operator and its Gaussian approximation from the linear mixing operator are shown to the right side panel. Clearly, the Gaussian approximation is adequate to approximate the MAX operator.

Notice that the above theorem doesn't guarantee the MAX operation on two *correlated* Gaussian random variables can also be approximated by a linear mixing operator. But this limitation will not affect the applicability of the theorem in STA. Correlated random variables modeling delay elements in a circuit timing graph can always be decomposed into independent delay elements according to Theorem 4 presented in section IV-B. And MAX on these correlated delay elements can be equivalent to a MAX on the decomposed delay elements followed by an ADD operator. So *only* MAX operation on independent random variables is involved for the purpose of STA.

IV. Variation Vector and Correlation Decomposition

The canonical timing model [6], [7], [10] is a powerful tool to represent the numerous delay elements for a given circuit. However, in its original format, it can only handle node delay correlations caused by global variations. In this work, we propose an *extended canonical timing model* that is capable of capture *all* the correlation between any pair of delay elements in the circuit be it a node delay or an edge delay.

Theorem 2 (Extended Canonical Timing Model): *Assume that there are N nodes and M global variations in the timing graph, if every node delay can be modeled by the canonical format of Equation (5), then every delay element, including all the node delays and edge delays will then have a extended canonical timing model as:*

$$X = \mu_X + \sum_{i=1}^{N} \alpha_{X,i} R_i + \sum_{j=1}^{M} \beta_{X,j} G_j \tag{12}$$

Proof: Using the mathematical induction principle:

Assertion I: If X is a node delay, then it will automatically have the extended canonical delay format because Equation (5) is a subset of Equation (12) in that for k^{th} node delay, only one $\alpha_{X,k}$ has non-zero value while all other $\alpha_{X,i \neq k}$ are set to zero.

Assertion II: If $X = A + B$ and delay elements A, B fit Equation (12) then X must have extended canonical delay format.

Assertion III: If $X = max(A, B)$ given that delay elements A, B fit Equation (12), Theorem 4 guarantees that $A = A' + W$, $B = B' + W$, $cov(A', B') = 0$ and A', B', W will also fit Equation (12). So $X = max(A, B) = max(A' + W, B' + W) = W + max(A', B') = W + \rho A' + (1-\rho)B' + \zeta$ according to Theorem 1. So X will have delay format of Equation (12).

Any delay element, if it is not a node delay, can ultimately be expressed as the result of one or multiple steps of ADD and/or MAX operations from node delays. So based on the above three assertions, the mathematical induction principle guarantees that all delay elements will have the extended canonical format of Equation (12). ■

A. Variation Vector

The extended canonical format of Equation (12) can be rewritten in a compacted vector format as

$$X = \mu_X + \boldsymbol{x}^T \boldsymbol{b} \tag{13}$$

where

$$\boldsymbol{b} \equiv [R_1, \cdots, R_N, G_1, \cdots, G_M]^T$$

is a random vector consisting of zero-mean, unity variance independent Gaussian random variables and

$$\boldsymbol{x} \equiv [\alpha_{X,1}, \cdots, \alpha_{X,N}, \beta_{X,1}, \cdots, \beta_{X,M}]^T$$

is a deterministic vector and is the *Variation Vector(v.v.)* of X.

So Each delay element(X) in a circuit will be uniquely represented by its mean(μ_X) and variation vector($\boldsymbol{x}$), noted as

$$X = X(\mu_X, \boldsymbol{x}) \tag{14}$$

With equation (12), 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 are represented by non-zero node sensitivity terms $\alpha_{X,k}$.

From definition, it is easy to verify the following properties for variation vector:

Theorem 3: *Assuming k and c are constants and $\boldsymbol{x}$, $\boldsymbol{y}$, $\boldsymbol{z}$ are variation vectors of delay elements X, Y, Z.*

(1) *X and $X + c$ have the same variation vector of $\boldsymbol{x}$;*
(2) *If $Z = X + Y$ then $\boldsymbol{z} = \boldsymbol{x} + \boldsymbol{y}$;*
(3) *If $Z = kX$, then $\boldsymbol{z} = k\boldsymbol{x}$.*
(4) $\sigma_X^2 = \boldsymbol{x}^T \cdot \boldsymbol{x} = ||\boldsymbol{x}||$
(5) $cov(X, Y) = \boldsymbol{x}^T \cdot \boldsymbol{y} = \boldsymbol{y}^T \cdot \boldsymbol{x}$

Property (1) indicates that variation vector remains unchanged if a constant is added to the corresponding random variable since variation vector contains only the variance information of the random variable while the added constant only affects the mean of the random variable. Properties (2) and (3) are the basis of variation vector propagation discussed later. Properties (4) and (5) make variation vector an convenient and systematic way to evaluate the variances and correlations for any delay elements.

B. Correlation Decomposition

Due to simplicity of handling independent delay elements, it is usually desirable to decompose correlated delay elements into independent ones. A typical method is to to use so called *Principle Component Analysis(PCA).* [12]

However, MAX operation is *not* communicative with general linear transformation operators (U):

$$max\{U(X,Y)\} \neq U(max\{X,Y\})$$

so little benefit can be obtained by applying it to calculate the MAX output for two correlated delay elements.

Based on the canonical timing model of delay elements, there exists a much more elegant way to decompose two correlated delay elements:

Theorem 4 (Correlation Decomposition): *Let delay elements X and Y in a circuit be represented in the extended canonical delay model representation. If X and Y are correlated, then there will be a third delay element W also in the extended canonical delay model representation such that $cov(X - W, Y - W) = 0$.*

Proof: Assume variation vectors of X and Y are $\boldsymbol{x} = (x_1, x_2, \cdots, x_{N+M})^T$ and $\boldsymbol{y} = (y_1, y_2, \cdots, y_{N+M})^T$, then a new variation vector of $\boldsymbol{w} = (w_1, w_2, \cdots, w_{N+M})^T$ can be constructed as:

$$w_i = min(x_i, y_i) \qquad i = 1, 2, \cdots, N + M \tag{15}$$

This construction procedure is illustrated in Figure 4

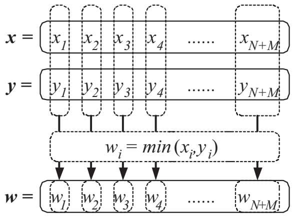


Fig. 4. Correlation Decomposition Procedure

So if a random variable W is defined as $W = W(\mu_W, \boldsymbol{w})$ with arbitrary mean value of μ_W, then $cov(X - W, Y - W) = (\boldsymbol{x} - \boldsymbol{w})^T(\boldsymbol{y} - \boldsymbol{w}) = 0$ since it is impossible for $(\boldsymbol{x} - \boldsymbol{w})$ and $(\boldsymbol{y} - \boldsymbol{w})$ to have common non-zero components in their variation vectors. ■

With this method of correlation decomposition, correlated delay elements X and Y are decomposed into X', Y' and W as $X = X' + W$ and $Y = Y' + W$ and $cov(X', Y') = 0$. What is more interesting is that this decomposition procedure is *communicative* with the MAX operation:

$$max(X' + W, Y' + W) = W + max(X', Y')$$

So all MAX operations on dependent delay elements can be simplified as a MAX operation on independent delay elements followed by an ADD operation and so that the computation will be greatly simplified.

V. Propagating Mean and Variation Vector

In a timing graph, the mean and variation vector of a node delay is obtained from technology extraction. To get orthogonality required by the mean and $v.v.$ representation of delay elements, Principle Component Analysis may be conducted after extraction.([10]) But this is done only once for a specific technology and so that is not considered as a part of STA. A STA algorithm, instead, will take those node's means and variation vectors as its input and calculate edge's mean and variation vector in the entire circuit.

A. Algorithm for ADD and MAX Operations

Through an ADD operation

$$Z(\mu_Z, \boldsymbol{z}) = X(\mu_X, \boldsymbol{x}) + Y(\mu_Y, \boldsymbol{y})$$

the mean and $v.v.$ propagation is straightforward:

$$\mu_Z = \mu_X + \mu_Y \tag{16}$$
$$\boldsymbol{z} = \boldsymbol{x} + \boldsymbol{y} \tag{17}$$

It is very easy to verify the consistency between this variation vector approach and Equation (2).

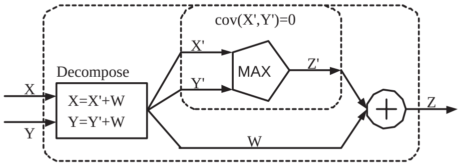


Fig. 5. $Z = max(X, Y)$ when $cov(X, Y) \neq 0$

The mean and variation vector propagation through MAX operation,

$$Z(\mu_Z, \boldsymbol{z}) = max\{X(\mu_X, \boldsymbol{x}), Y(\mu_Y, \boldsymbol{y})\}$$

is illustrated in Figure 5 where totally four computation steps are involved:

(1) Correlation Decomposition

$$X(\mu_X, \boldsymbol{x}) = X'(\mu_{X'}, \boldsymbol{x'}) + W(\mu_W, \boldsymbol{w})$$
$$Y(\mu_Y, \boldsymbol{y}) = Y'(\mu_{Y'}, \boldsymbol{y'}) + W(\mu_W, \boldsymbol{w})$$

where $cov(X', Y') = \boldsymbol{x'}^T \cdot \boldsymbol{y'} = 0$

(2) Calculate $\mu_{Z'}$ and $\sigma_{Z'}$ for $Z' = max(X', Y')$ from Clark Equations (3 and 4)

(3) Calculate contribution factor ρ for $Z' = max(X', Y')$ from Equation (11) of MAX linearization Theorem.

(4) Final Results for $Z = max(X, Y)$

$$\mu_Z = \mu_{Z'} + \mu_W \tag{18}$$
$$\boldsymbol{z} = \rho\boldsymbol{x'} + (1 - \rho)\boldsymbol{y'} + \boldsymbol{w} \tag{19}$$

If there are more than two delay elements involved in the MAX operation, then MAX is done iteratively by MAX two delay elements at each iteration.

B. Exploration of Sparsity

Since there are N nodes in a timing graph, and each node delay as well as corresponding edge delay will have a $N+M$ dimensional variation vector, the total computation and storage required will be $O(N(N+M)) \approx O(N^2)$. However, while working with benchmark circuits, we noticed that many components in the variation vector have very small values, indicating that their contributions to the overall correlation evaluation is insignificant. By setting these small coefficient to zero, the variation vector will become a sparse vector that contains many zero components.

Motivated by this observation, we developed a novel technique called the *flexible vector format* to exploit the sparsity of the variation vector. In particular, we focus on curtailing node part of the variation vector, $\alpha_{X,i}$, that have small values. For this purpose, a *drop threshold* is selected so that if $\alpha_{X,i}$ is smaller than this threshold, it is deemed to have small value and will be placed into a drop candidate pool to be pruned from the variation vector representation.

However, dropping $\alpha_{X,i}$ with small magnitude is the same as applying truncation to the variation vector. In subsequent computations, the quantization error may accumulate, causing non-negligible error. This is a problem that can not be overlooked for large circuits. Our solution to this problem is to lump those components in the drop candidate pool into a single correction term

$$x_{pool} \quad = \quad \sqrt{\sum x^2_{\textit{dropped Components}}} \tag{20}$$

When two variation vectors merge through either ADD or MAX operation, their pooling components are assumed to be independent. Hence,

$$z_{pool}(ADD) \quad = \quad \sqrt{x^2_{pool} + y^2_{pool}} \tag{21}$$

$$z_{pool}(MAX) \quad = \quad \sqrt{\rho^2 x^2_{pool} + (1-\rho)^2 y^2_{pool}} \tag{22}$$

C. Complexity and Path Correlation Length

Using this drop and pool mechanism, what is really dropped during computation is then the path correlation. So the length of the variation vector actually gives a good indication to the extent of path correlation in the circuit. The *path correlation length*(Γ) of the circuit is then defined to be the average length of node part of the pruned variation vectors for a given drop threshold.

With this notation, the computation complexity can be reduced from $O(N^2)$ down to $O[(\Gamma + M) \cdot N]$. Simulation results indicated that $\Gamma << N$ and is not a function of N. Hence it represents a significant reduction of computation and storage.

VI. Simulation Results and Discussions

The above described algorithm has already been implemented in C/C++ with name of **AMECT** and tested by ISCAS85 benchmark circuits.

Before testing, however, all benchmark circuits are remapped into a library which has gates of *not, nand2, nand3, nor2, nor3 and xor/xnor*. Table I summarizes the gate count for each test circuit after gate re-mapping.

TABLE I
ISCAS85 Benchmark Circuits

Name	c432	c499	c880	c1335	c1908
Gate Counts	280	373	641	717	1188
Name	c2670	c3540	c5315	c6288	c7552
Gate Counts	2004	2485	3865	2704	5355

All library gates are implemented in $0.18\mu m$ technology and their delays are characterized by Monte Carlo simulation with Cadence tools assuming all variation sources, either process variations or operational variations, follow Gaussian distribution.

For illustration purpose, only three parameter variation are considered global: channel length(L), supply voltage(Vdd) and temperature(T). All other variation sources, specified in the $0.18\mu m$ technology file, are assumed to be localized in the considered gate only. Also we don't address the spatial dependency of the gate delays just for demonstration purpose. In real life, gate timing parameters are position dependent and our method is still applicable.

A. Accuracy and Performance

Monte Carlo simulation results with 10,000 repetitions are used as "Golden Value" for each benchmark circuit. Each repetition is a process of static timing analysis by fixing global and node variation into a set of randomly sampled values. The global variations are sampled once for each repetition while node variation for each gate is newly sampled every time when the gate is computed.

Table II summarizes the edge delay distribution parameters at the primary output of each testing circuit from Monte Carlo(M.C.) and two flavors of STA methods using **AMECT**. For comparison purpose, a fourth STA method *NoCorr* with no correlation considered is also implemented and simulated. μ and σ are mean and standard variation of the distribution. $\tau_{97} = \mu + 2\sigma$ is the delay estimation at confidence level of 97%. The accuracy of STA methods compared with Monte Carlo method, is evaluated in Table III.

The drop threshold in **AMECT** will determine the extent at which the path correlation is considered. Method *HighAccu* is the high accuracy version of **AMECT** when drop threshold is set into 1% and most path correlations are considered while method *HighPerf* is the high performance version of **AMECT** with drop threshold of 100% and only global correlation is considered.

From Table III, it is very clear that method *NoCorr* fails to give reasonable variance estimation because no correlation is considered which gives a good example for the importance of correlations in STA. It is also notable that *NoCorr* can still have fairy reasonable mean estimation which tells that the mean delay is not so sensitive to the correlation. This interesting phenomenon may come from the ADD operation whose variance is very sensitive to the input correlation.

TABLE II

TESTING RESULTS FOR ISCAS BENCHMARKS

Circuit	STA Method	CPU Time[s]	Delay Distribution[ps]		
			μ	σ	τ_{97}
c432	M.C.	6.449	1288.8	219.3	1727.5
	HighAccu	0.030	1299.0	220.4	1739.9
	HighPerf	0.010	1348.6	216.0	1780.7
	NoCorr	0.000	1392.6	22.3	1437.1
c499	M.C.	8.182	1073.6	178.9	1431.4
	HighAccu	0.030	1084.8	180.5	1445.8
	HighPerf	0.010	1125.4	178.3	1482.0
	NoCorr	0.000	1148.6	20.5	1189.5
c880	M.C.	14.831	1445.4	266.3	1977.9
	HighAccu	0.050	1447.6	264.9	1977.3
	HighPerf	0.010	1463.1	264.2	1911.5
	NoCorr	0.010	1471.6	16.5	1504.5
c1355	M.C.	19.007	1445.4	251.4	1948.3
	HighAccu	0.071	1460.9	250.7	1962.3
	HighPerf	0.010	1529.4	249.0	2027.4
	NoCorr	0.000	1536.8	12.5	1561.8
c1908	M.C.	35.801	1828.2	327.3	2482.8
	HighAccu	0.150	1841.9	328.4	2498.6
	HighPerf	0.030	1881.7	326.5	2534.7
	NoCorr	0.010	1895.1	27.0	1949.2
c2670	M.C.	72.163	2097.0	382.9	2862.8
	HighAccu	0.181	2104.4	382.9	2870.1
	HighPerf	0.050	2161.8	379.7	2921.2
	NoCorr	0.020	2193.1	22.0	2237.1
c3540	M.C.	84.020	2747.2	498.8	3744.8
	HighAccu	0.240	2752.3	502.1	3756.5
	HighPerf	0.050	2850.3	500.6	3851.5
	NoCorr	0.020	2859.7	23.4	2906.5
c5315	M.C.	140.832	2399.3	441.7	3282.6
	HighAccu	0.641	2404.8	442.2	3289.2
	HighPerf	0.080	2474.1	441.3	3356.7
	NoCorr	0.040	2544.8	31.7	2608.2
c6288	M.C.	114.235	6740.1	1286.8	9313.6
	HighAccu	5.198	6775.9	1275.1	9326.2
	HighPerf	0.070	7290.8	1273.1	9836.9
	NoCorr	0.030	7325.1	14.9	7355.0
c7552	M.C.	202.972	1911.7	348.7	2609.0
	HighAccu	0.571	1916.6	353.8	2624.2
	HighPerf	0.110	1974.3	352.5	2679.3
	NoCorr	0.050	2027.4	26.3	2080.1

TABLE III

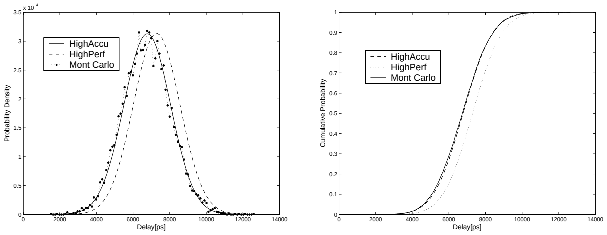
DISTRIBUTION ERROR RESPECTING TO MONTE CARLO RESULTS

Circuit	Mean Error($\delta\mu$)			Variance Error($\delta\sigma$)		
	HighAccu	HighPerf	NoCorr	HighAccu	HighPerf	NoCorr
c432	0.79%	4.64%	8.05%	0.50%	1.50%	89.8%
c499	1.04%	4.82%	6.99%	0.89%	0.34%	88.5%
c880	0.15%	1.22%	1.81%	0.53%	0.79%	93.8%
c1355	1.07%	5.81%	6.32%	0.28%	0.95%	95.0%
c1908	0.75%	2.93%	3.66%	0.27%	0.24%	91.8%
c2670	0.35%	3.09%	4.58%	0.00%	0.84%	94.3%
c3540	0.19%	3.75%	4.10%	0.66%	0.36%	95.3%
c5315	0.23%	3.12%	6.06%	0.11%	0.09%	92.8%
c6288	0.53%	8.17%	8.68%	0.65%	1.06%	98.8%
c7552	0.25%	3.27%	6.05%	1.46%	1.09%	92.5%

Table III also shows that *HighPerf* have significantly larger error in mean estimation than *HighAccu*. This is reasonable because *HighAccu* will overestimate the mean at every MAX operation due to smaller correlation considered and this over estimation is accumulated through distribution propagation. It is also interesting to notice that *HighPerf* and *HighAccu* give similar accuracy in variance estimation. This is possibly because of the fact that the variance is dominated by global variation in the tested cases.

Of course, Monte Carlo simulation gives the best STA results but with big runtime penalty. **AMECT** runs order-of-magnitude faster but can provide both mean and variance estimation almost as accurate as Monte Carlo does if most of the path correlations are considered as the cases of *HighAccu* shown in Table II and III. If the accuracy on mean estimation can be relaxed, then the drop threshold can be higher and **AMECT** will give some mean overestimation but with better performance in runtime. In another word, **AMECT** is parameterized by the drop threshold and can be used to trade-off accuracy with performance in some circumstances.

To further elaborate the accuracy of **AMECT**, Figure 6 shows the $p.d.f.$ and $c.d.f.$ for circuit c6288 from three methods: Mont Carlo and two methods of **AMECT**(*HighAccu* and *HighPerf*). Apparently enough, **AMECT** shows excellent accuracy if most path correlation is considered as in method *HighAccu*.



(a) $p.d.f.$ from Three Methods (b) $c.d.f.$ from Three Methods

Fig. 6. $p.d.f.$ and $c.d.f.$ comparison for c6288 from three methods

TABLE IV

PATH CORRELATION LENGTH AT 1% OF DROP THRESHOLD

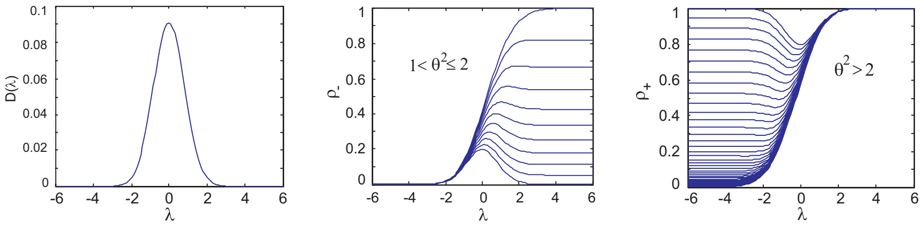
Name	c432	c499	c880	c1335	c1908
Γ	22.0	11.1	14.2	19.3	27.0
Name	c2670	c3540	c5315	c6288	c7552
Γ	15.4	21.2	14.4	80.9	16.0

B. Path Correlation Length

It has been mentioned in Section V-C that path correlation length (Γ) provided by **AMECT** is an interesting macro property of the simulated circuit and gives a good indication of the extent of the path correlation existing in that circuit. For the above ISCAS circuits, the path correlation length(Γ) at drop threshold of 1% is summarized in Table IV.

From Tale IV, we can firstly conclude that the correlation length Γ is much smaller than the circuit size and basically independent on the circuit size since it remains about $10-20$ when circuit size changes dramatically. This observation helps the conclusion we made before about the complexity reduction of **AMECT** by using the technique of flexible vector format.

Secondly, the only exceptional high path correlation length among the tested circuits happens with the circuit c6288 which is known as a 16-bit array multiplier. Since there



(a) Verifying Inequality 23 (b) Verifying $0 < \rho_- < 1$ (c) Verifying $0 < \rho_+ < 1$

Fig. 7. Numerical Results for Appendix I and II

are large amount of equal delay paths in the circuit, large path correlation length is natural: Few node variation can be dropped due to the equal importance.

VII. Conclusions

This paper presents a novel method for block-based statistical timing analysis. Applying the generally accepted Gaussian assumption, we firstly disclose that the MAX operation can be approximated by linear supposition of its inputs. Secondly we extend the commonly used canonical timing model into a vectorized format, variation vector. We also disclose a novel method to decompose correlated timing variables into independent ones to simplify computation. With these theoretical progress, we are able to evaluate and propagate the global and path correlation systematically in the circuit timing graph.

We also design a novel algorithm, **AMECT** which treat both global and path correlation simultaneously and systematically. This algorithm, with the help with a new flexible vector format achieves high accuracy and high performance at the same time as tested by ISCAS circuits and compared with Monte Carlo "golden value".

Appendix I
Existence of Contribution Factor

Clearly, $max(X + c, Y + c) = c + max(X, Y)$ and $max(kX, kY) = k \cdot max(X, Y)$ for any constant $k > 0$ and c. So the variance matching Equation (9) will not change if X, Y and Z are shifted and positively scaled at the same time. So the solution of ρ to the scaled and/or shifted equation will also be the solution to the original one. So no generality will lose if we additionally assume $\mu_X = \mu$, $\mu_Y = 0$, $\sigma_X^2 = \sigma^2$, and $\sigma_Y^2 = 1$.

Applying results from Equations (3 and 4), $E(Z) = \mu Q + \theta P$, $E(Z^2) = (\mu^2 + \sigma^2 - 1)Q + \mu\theta P + 1$, $\theta^2 = 1 + \sigma^2$ and $\lambda = \mu/\theta$, It is then sufficient to prove the inequality (10) if:

$$D(\lambda) = A - Q^2 \geq 0 \tag{23}$$

where $A = (1 + \lambda^2 - 2\lambda P)Q + (\lambda - P)P - \lambda^2 Q^2$.

Since $D(\lambda)$ is only a function of λ, numerical evaluation in Figure 7(a) shows that it is always positive. So the inequality (10) is proved and the existence of the contribution factor is guaranteed.

Appendix II
Bounding the Contribution Factor

After scaling and shifting described in Appendix I, the root ρ will have the form of:

$$\rho_\pm = \frac{1}{\theta^2}[1 \pm \sqrt{1 - 2Q\theta^2 + A\theta^4}] \tag{24}$$

If $\sigma_Y^2 \geq \sigma_X^2$, then $1 < \theta^2 \leq 2$ and $0 < \rho_- < 1$ as shown in Figure 7(b). Similarly, if $\sigma_Y^2 < \sigma_X^2$, then $\theta^2 > 2$ and $0 < \rho_+ < 1$ as shown in Figure 7(c).

So by switching the contribution factor ρ between ρ_- and ρ_+ according to the relative magnitude of σ_X^2 and σ_Y^2, $0 < \rho < 1$ can always be satisfied.

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," *Design Automation Conference, 2002. Proceedings. 39th*, pp. 566 – 569, June 2002.

[2] M. Orshansky, "Fast computation of circuit delay probability distribution for timing graphs with arbitary node correlation," *TAU'04*, Feb 2004.

[3] M. Orshansky and K. Keutzer, "A general probabilistic framework for worst case timing analysis," *Design Automation Conference, 2002. Proceedings. 39th*, pp. 556 – 561, June 2002.

[4] A. Agarwal, D. Blaauw, V. Zolotov, S. Sundareswaran, M. Zhao, K. Gala, and R. Panda, "Statistical delay computation considering spatial correlations," *Design Automation Conference, 2003. Proceedings of the ASP-DAC 2003. Asia and South Pacific*, pp. 271 – 276, Jan 2003.

[5] A. Agarwal, V. Zolotov, and D. Blaauw, "Statistical timing analysis using bounds and selective enumeration," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 22, no. 9, pp. 1243 –1260, Sept 2003.

[6] C. Visweswariah, K. Ravindran, and K. Kalafala, "First-order parameterized block-based statistical timing analysis," *TAU'04*, Feb 2004.

[7] A. Agarwal, D. Blaauw, and V. Zolotov, "Statistical timing analysis for intra-die process variations with spatial correlations," *Computer Aided Design, 2003 International Conference on. ICCAD-2003*, pp. 900 – 907, Nov 2003.

[8] S. Bhardwaj, S. B. Vrudhula, and D. Blaauw, "τau: Timing analysis under uncertainty," *ICCAD'03*, pp. 615–620, Nov 2003.

[9] A. Devgan and C. Kashyap, "Block-based static timing analysis with uncertainty," *ICCAD'03*, pp. 607–614, Nov 2003.

[10] H. Chang and S. S. Sapatnekar, "Statistical timing analysis considering spatial correlations using a single pert-like traversal," *ICCAD'03*, pp. 621–625, Nov 2003.

[11] C. Clark, "The greatest of a finite set of random variables," *Operations Research*, pp. 145–162, March 1961.

[12] D. F. Morrison, *Multivariate Statistical Methods*. NewYork, McGraw-Hill, 1976.