A Survey of SSTA Techniques with Focus on Accuracy and Speed

Bhaghath P J
Department of Electronics and Communication Engineering
Amrita Vishwa Vidyapeetham University, Coimbatore, India

Ramesh S R
Department of Electronics and Communication Engineering
Amrita Vishwa Vidyapeetham University, Coimbatore, India

ABSTRACT

Timing analysis plays a vital role in chip design, which analyze whether a chip design meets the timing constraints. The main objectives of timing analysis are speed and accuracy. There are two engines for timing analysis namely Statistical Timing Analysis (STA) and Statistical Static Timing Analysis (SSTA). VLSI CAD has been gaining a lot of interest in both STA and SSTA. As technology continues to advance deeper in to the nanometer regime, a tight control on the process parameters is increasingly difficult. To account these process parameters which are probabilistic in nature while performing timing analysis SSTA is preferred. The main goal of SSTA is to improve the accuracy without any reduction in speed by considering process variations. This paper presents a survey of SSTA approaches and techniques for improving accuracy and speed by considering the topological correlations and spatial correlations.

General Terms

Timing analysis, STA, SSTA, Speed, Accuracy

Keywords

VLSI CAD, Arrival Time, Required Arrival Time, Slack, Critical path, Conditional criticality, Complementary Slack, Arrival tightness probability, Ellipse graph.

1. INTRODUCTION

On Chip Variation (OCV) increases when going in to nanometer regime especially from 130nm onwards [15]. These OCVs are of two types. They are random and systematic. In digital circuits, the time at which a signal arrives at a destination point is affected by several factors termed as variations. These variations may be temperature, voltage and process variations. The purpose of timing analysis is to ensure whether the signal reaches its destination as per the timing constraints. The goal of timing analysis is that despite all possible variations it should ensure proper circuit operations by assuring that the signals arrive neither too early nor too late. Earlier method chosen for this key purpose was static timing analysis. The current trend in IC designing is to reducing the size as much as possible. With CMOS technology scaling down to nanometer regime process variations has been increased. This seriously affects the interconnect delay; gate delay etc. STA is deterministic in nature. Since process variations cannot be accounted the only choice is SSTA. Fujitsu laboratories [11] started research on SSTA in 2003 and has applied SSTA technologies to processor and ASIC designs.

VLSI CAD approaches both STA and SSTA based on graph theory concepts with help of C/C++ compilers. It uses basic concepts like Directed Acyclic Graph (DAG) and Augmented Graph. If virtual node and virtual edges are added to the primary input nodes of a graph, is termed as source node. Similarly if a virtual edge and virtual node are added to the primary output nodes of graph, is referred as sink node and the resultant graph is known as augmented graph. If a graph has no cycles or feedback and also unidirectional known as Directed Acyclic Graph (DAG).

1.1 STA

Static timing analysis [14] is one of the engines widely adopted for timing analysis. It validates timing performance of a digital circuit by considering all possible paths for timing violations under worst case conditions. The reasons behind its popularity is linear runtime with circuit size, conservative which means it overestimate the delay of longest path and underestimate the delay of shortest path, effectively addressing paths like false path, multiple cycle path etc.

It assumes that the process parameters like temperature, voltage and device parameters like oxide thickness, gate length are fixed. When it comes to nanometer regime these parameters vary randomly. Due to the deterministic characteristics of STA, it is unable to incorporate them in the analysis. In addition to this for each parameter the STA algorithm has to be run individually, thereby obtain multiple corner files are obtained. As the size of corner file increases, it is difficult to manage.

1.2 SSTA

To incorporate process variations in timing analysis, SSTA is preferred [14] rather than STA. The ability to consider parameters which are random in nature make it more popular among all engines or tools like statistical Hspice , Monte carlo analysis etc The strength of SSTA is ability to perform delay calculations like Arrival time, slack etc by considering process parameters in timing analysis.

1.3 Sources of Process Variations

Steps of design process and their resulting timing uncertainties is depicted in fig:1 [14]. As the device size shrinks the device size the process parameters have predominant role in timing violations. The random behaviour of process parameters makes the optimization a difficult task, thereby reducing the accuracy.

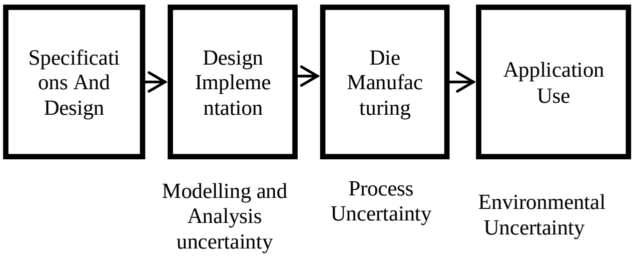
Specificati ons And Design
Design Impleme ntation
Die Manufac turing
Application Use
Modelling and Analysis uncertainty
Process Uncertainty
Environmental Uncertainty

Fig 1: Steps of Design Process and Their Resulting Timing Uncertainties

The errors that occur while performing timing analysis can be classified in to three main categories.

1. Modelling and analysis errors-
2. Manufacturing variations
3. Operating context variations

Once the design specifications are ready, next step is to model the design. After modeling, it undergoes several testing and verification stages like power consumption, delay, layout, floor planning etc. The outcome of this step may deviate from the expected ones. After making necessary corrections, it goes to fabrication level.

The challenges faced at this level are variations due to any limitation in process, process equipment imperfections and imprecisions etc. After fabrication, the fabricated device ready to use. There it faces uncertainties from environment.

1.4 Sources of Physical Variations

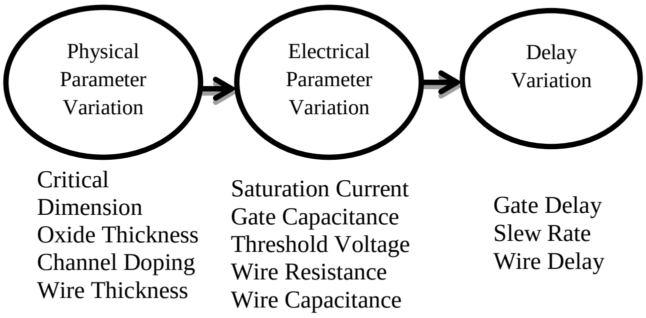

Fig 2: Sources of Variations in SSTA

Sources of Variations in SSTA is shown in fig: 2 [14]. Any variations in physical parameters like critical dimension, oxide thickness etc., will lead to electrical variations. Those electrical variations cause delay variations like gate delay, wire delay, slew delay.

1.5 Classification of Process Parameters

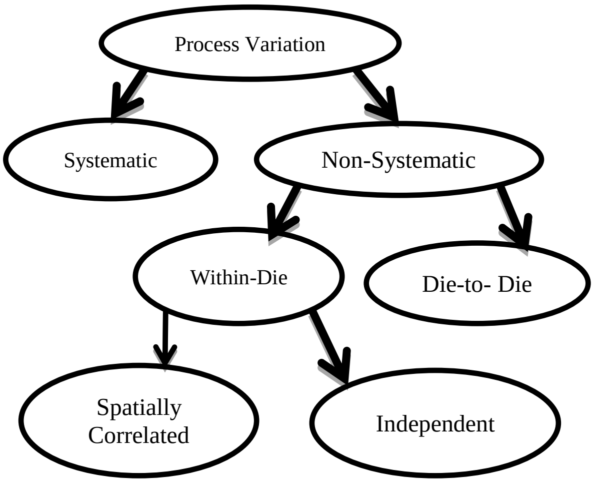

Fig 3: Classification of Physical Parameter Variation

Process variations are mainly classified in to two as shown in fig:3[14]. There are systematic and non-systematic variations. Systematic variations can determine earlier stages of design, but process parameters which are random in nature come under non-systematic. These non-systematic variations are classified as within-Die and Die to Die. Die to Die variations are also called as global variation and within Die variations as local variations. In Within-Die variations, the process parameters affect the devices which are close to each other in the same amount. This will cause correlations between them called as spatial correlation and each devices affected independently by these parameters are called independent or local variation.

1.6 Challenges in SSTA

1) Topological correlation: This is mainly due to path reconvergence. Path reconvergence means, a path starting from common node gets separated in to multiple paths again these paths converge to another common node. These mainly affect the Max operator.

2) Spatial correlation: Correlation exists between the devices if they share common boundaries of grids. This challenge is captured and propagated during SSTA thereby increasing the accuracy. However, it increases the complexity of algorithms and affects both Max and Sum operator.

1.7 Approaches in SSTA

Path based technique: It finds the delay of each path. At the sink node it takes the maximum of those path delays. The advantage of this approach is it will not miss any critical path. The limitation of this approach is no slandered method to choose right path for analysis. Also the runtime increases exponentially.

Block based analysis: In this approach it takes each interconnect and components as block. At each block it applies MAX operator to find the arrival time if those have multiple inputs. The runtime is linear and progressive computation is carried out. Due to non-linear behaviour of MAX operator, there by reduction in accuracy. Block based analysis uses wide number of MAX operator.

2. KEY DEFINITIONS

Arrival Time: It is the time required to take a signal to arrive at a certain point in a circuit.

Required time: It is the latest time at which a signal can arrive without making the clock cycle longer than desired.

Slack: It is defined as the difference between arrival time and the required time.

Critical path: it is the longest path between input and output.

Criticality of a path: It is the probability that manufacturing a chip in which this path is critical.

Criticality of set of paths: It is the probability of manufacturing a chip in which at least one from this set is critical.

Criticality of an edge (node): It is the probability of manufacturing a chip in which this edge (node) is on the critical path.

Edge slack: It is the maximum delay of all paths going through the edge.

Complement Edge slack: It is the maximum delay of all paths not going through the edge.

3. DIFFERENT TECHNIQUES OF SSTA AND ITS CHALLENGES

3.1 PERT like Traversal

Program Evaluation and Revaluation Technique (PERT) [2] starts by converting netlist file into augmented graph. This method finds the critical paths by traversing the graph in topological order using basic algorithms. In this the delays are converted in to probability density function. To make the analysis easier, Principal Component Analysis (PCA) is used. It treats the correlated parameters as uncorrelated set. By using max operation and sum operator, the arrival time is found out. Even though the computation of complexity of this analysis is linear; it depends on the circuit size. Also accuracy, runtime depends on the same. Accuracy reduction also occurs due to repeated number of literals.

3.2 Block Based Analysis with Uncertainty

Block based analysis is used to avoid the difficulty in choosing the right critical path of interest. Here the process parameters as treated as random in nature. Interconnect delays are modeled as PDF and gate delays are modeled as CDF [3], thereby it consider the uncertainty behaviour of process parameters, to reduce computational complexity used piecewise linear approach .Dependency list is maintained to consider about reconevrgent fanout. But as circuit size increases. it is very difficult to maintain dependency list , thereby it requires higher memory capacity.

3.3 Criticality Computation in Parameterized Statistical Timing

A path with longest delay is called as critical path. A circuit will have large number of critical path. Among the critical path, our choice is to determine which one is more critical. The basic idea behind this method is suppose there are two paths P1 and P2. These dies are generally called as process subspace. Criticality of path depends on the probability manufacturing a chip in this process subspace. Similarly criticality of edge (node), criticality of set of paths can be found out. To find this traversing the graph is traversed in the topological order using BFS algorithm. Approaches and technique used for this purpose is tightness probability and cutset computation [6]. To make the optimization simpler, the path which is critical and which causes a timing violation is chosen. Conditional probability of manufacturing a chip in which this path (edge, node) is critical, conditional upon the chip violating its timing constraints. Cutset approach effectively calculates the above mentioned factors, but it is not straight forward. Also it is non-incremental and computational complexity increases as size increases.

3.4 First order Incremental Block Based Statistical Timing Analysis

In this approach canonical forms are used to effectively address the local variation and global variation. A graph is traversed in topological order, the local variations treated as root of sum of squares, which reduces the spread of long path which consisting of main stages [5]. For finding criticality, graph traversal is done in both forward and reverse order based on few properties. The concept used here is arrival tightness probability (ATP). Once a small change is made in the circuit, it is necessary to get answers for queries about a particular point or device. This requirement is satisfied by new approach called level limiting. In order to speed up the above procedure another concept is added to the above one called as dynamic binding. This method considers only first order parameters. If this approach is extended for second order parameters the outcome will be more accurate. Also the dependency between criticality probabilities is not considered. Once changes are made to consider the above limitation this approach will be an error less and more accurate. However the run time of this approach and excessive usage of MAX operator is to be considered since it is a block based analysis.

3.5 Lump method and Analytical Spatial Correlation

Existing method for considering spatial correlation is Quad tree model. Reduction in accuracy is a major drawback since it is fails to give similar correlation and thereby reduction in accuracy is. The new approach introduced here is fanout pruning method. This method, finds out the path re-convergence of each component. At that path reconevrgent node, lump the local sensitivity a parameters, thereby reducing redundancy [7]. It increases the accuracy. By making grid size proportional to spatial correlation distance the limitation of quad tree model can be overcome. This technique is known as analytical spatial correlation. By using above methods the accuracy increases with linear complexity. As the circuit size increases, because of lump method it introduces quantization error.

3.6 CLECT

Conventional canonical expression effectively addresses the global variation but not local variations. Conditional linear MAX/ MIN Approximation and Extended Canonical timing Model (CLECT) [4] makes some modification to that expression, there by considering each devices local variation till the end and it is called as Extended Canonical Model [4]. It increases the accuracy. One limitation of block based analysis is the wide usage of MAX operator, because even if the inputs to a MAX operator is Gaussian the output may be a non-Gaussian function. This will reduce the accuracy. To increase the accuracy a new term called MAXTUPLE introduced is intr. Each time it checks the linearity behaviour of MAX operator for the corresponding inputs by analytical method. If the result is linear then MAX operator applies. Otherwise application of MAX/MIN is postponed temporarily with a new term called MAXTUPLE. All parameters does not have significant role in timing violation. In such cases, dropping such parameters does not have that much impact on accuracy. Such parameters sensitivity coefficients are set to zero. If this approach is performed for large circuits it may accumulate the quantization error. To overcome this issues a new approach called Drop Lump method [4] with the help of threshold value is used. By using above mentioned approaches overall accuracy increases but each approach has its own limitations. For example, skewness of Max operator is determined by analytical method. It increases the complexity. Also computational complexity for ECM is O (N^2) instead of O (N) [7]. Drop and lump depends on threshold value. To fix that threshold value, a perfect method is needed. If the threshold value chosen is wrong, it may lead to timing violation.

3.7 Clustering based Cutset Pruning Ordering Method

In conventional cutset based approach if an edge goes through two or more cutset, it increases the runtime. If an edge passes through more than one cutset it is named as MC-edge. First method is used zone based computation. This method contributes to the global errors in other edges in the cutset. In addition to that MAX operator also contributes to accuracy reduction. To overcome the limitation of cutset based approach, a new method named Clustering based cutset pruning ordering method is proposed [9]. In this method, new terms are introduced. They are dominant and non-dominant edges. This is based on threshold value. If the local criticality is above this threshold values it is dominant otherwise it is non-dominant. In order to avoid contribution to global errors as in zone based approach, it removes the non-dominant edges from the cutset there by their contribution to reduction in accuracy reduces. However this method depends on threshold value. Hence efficiency of this method depends on capability of a method for choosing that threshold value. Also complexity for cutset pruning is not linear. The non-linear computational complexity of the above mentioned techniques is reduced by clustering based pruning and ordering. There is a wide usage of MAX operator, thereby accuracy reduces.

3.8 Incremental Criticality and Yield Gradients

In this approach two important analysis measures of SSTA named as criticality and yield gradients are found effectively after making some modifications to circuit [12]. Conventional cutset is basically non-incremental in nature. In cutset after modification whole analysis has to start from the beginning. It will increase the time and cost. To overcome this two methods are introduced. They are instrumentality via probability identity and instrumentality via reconstruction of complement stack based on cutset. Both approaches avoid need for analysis from beginning after modification. There by they reduces the time for analysis. Also it addresses incremental yield gradients. To reduce the runtime for calculating chip slack after modification, the edges to source node are reconstructed into tree like structure. Assigning zero delay for the edges in tree structure, by making an assumption that there is no delay to propagate the effect after any modification in any part of the chip. This requires more number of MAX operators than exactly needed.

3.9 Incremental SSTA with Gate Timing Yield Emphasis

If more number of modifications are given to a local spot it is very difficult to compute criticality and yield gradients with conventional technique. To overcome this a new approach called timing yield emphasis is introduced here [13]. This approach depends on incremental threshold. If the timing yield of particular gate is smaller than the threshold value, then delay update is skipped. It makes the timing yield of circuit as a function of timing yield of gate. Also it finds the efficiency and error in each method.

3.10 Fitting Spased Approach

A new technique [16] developed to overcome the limitation of PCA based approach is fitting spased approach. Computational complexity and making SSTA computation more complex are the limitations of PCA based approach. Because of artificial transformation of timing parameters, it is unable to determine which area of chip is responsible for timing violation. In fitting based approach, is defined a one dedicated spatial random variable for each grid which is independent of all other spatial random variables. This is based on new term called correlation index 'I'. This model sticks on the basic concept that two random variables have the same correlation if they have same distance.

3.11 Sparse Matrix based SSTA

It is based on the path gate incidence matrix. Path delay obtained by multiplying the delay vector with sparse matrix. Circuit delay is the maximum of all path delay. This approach is well defined for both STA and SSTA [17]. The merit of this technique is highlighted in such a way that no restriction is imposed on process parameter distribution. This method considers slope propagation by incorporating polynomial based delay model. As the circuit size increases runtime and computational complexity increases

3.12 Refactoring Technique

The arrival time for each vertex is calculated in topological order and is depicted in fig: 4 [19]. A Max-Plus-Expression (MPE) at the sink node before and after refactoring technique is represented by equation 3 and equation 6. From those equations it is clear that, the number of repeated literals is reduced from eight to six. The proposed method [19] is to overcome the reduction in accuracy due to repeated number of literals.

$$At(v1) = a.b \tag{1}$$

$$At(v2) = c.(d + a.b) \tag{2}$$

Before refactoring technique

$$At(v3) = \Big(e + \big(c.\big(d + (a.b)\big)\big)\Big).(f + (a.b)) \tag{3}$$

After refactoring technique

$$At(v3) = \Big(e + \big(c.\big(d + (a.b)\big)\big)\Big).\big(f + (a.b)\big) \tag{4}$$

$$= (e + c).\big(e + d + (a.b)\big).\big(f + (a.b)\big) \tag{5}$$

$$= (e + c).\big(\big((e + d).f\big) + (a.b)\big) \tag{6}$$

Arrival time of each vertex calculated in the topological order is shown in fig: 4. The basic idea is each literal represents a process parameter. If a literal repeated multiple times, it will definitely lead to reduction in accuracy. For this purpose the concept of ellipse graph and division operation is well defined. The concept of refactoring technique is addition operation is distributive in nature. Based on this Division operation are defined. Division operation is carried out based on few properties. If the graph follows which follows those properties it is termed as ellipse graph. There are two refactoring technique mainly Static and Dynamic.

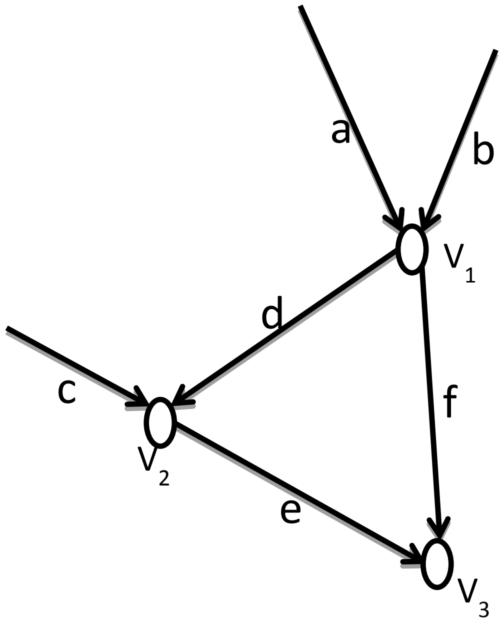

Fig 4: Arrival Time of Each Vertex Calculated in the Topological Order

Static method assumes that redundant literal causes same amount of error whereas in dynamic each redundant literal causes different amount of error. Hence the dynamic method determines the upper bound and lower bound of delay and increases accuracy.

3.13 Criticality Computation using Unary and Conditional Operator

The proposed method is to overcome the error due to usage of MAX operator while computing criticality [20]. Error occurs due to non-linearity behaviour of max operator. This is overcome by using two new operators named Unary and Conditional operators. Conditional operator well defined in [18]. Based on few properties defined an algorithm to find out the criticality of path or node or edge. The proposed method is well defined SSTA that both with and without refactoring technique [19], thereby increase in accuracy. The criticality of node is computed from those of its outgoing edges by union operator and that of edge by conditional operator.

4. FUTURE SCOPE

This survey presented the various issues on timing analysis. Reconvergent paths which arise due to topological correlation degrade the performance of timing graphs. From this survey, a general conclusion can be drawn that no effective method to consider path reconvergence is addressed in the literature. If there exists any method to effectively address this challenge, it will become a significant achievement in timing analysis. Incorporating fuzzy with SSTA instead of probabilistic approach will make timing analysis more productive.

5. REFERENCES

[1] Clark, Charles E. "The greatest of a finite set of random variables." *Operations Research* 9, no. 2 (1961): 145-162.

[2] Chang, Hongliang, and Sachin S. Sapatnekar. "Statistical timing analysis considering spatial correlations using a single PERT-like traversal." In *Proceedings of the 2003 IEEE/ACM international c onference on Computer-aided design*, p. 621. IEEE Computer Society, 2003.

[3] Devgan, Anirudh, and Chandramouli Kashyap. "Block-based static timing analysis with uncertainty." In *Proceedings of the 2003 IEEE/ACM international conference on computer-aided design*, p. 607. IEEE Computer Society, 2003.

[4] Zhang, Lizheng, Weijen Chen, Yuhen Hu, and CC-P. Chen. "Statistical timing analysis with extended pseudo-canonical timing model." In *Design, Automation and Test in Europe, 2005. Proceedings*, pp. 952-957. IEEE, 2005.

[5] Visweswariah, Chandramouli, Kaushik Ravindran, Kerim Kalafala, Steven G. Walker, Sambasivan Narayan, Daniel K. Beece, Jeff Piaget, Natesan Venkateswaran, and Jeffrey G. Hemmett. "First-order incremental block-based statistical timing analysis." *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on* 25, no. 10 (2006): 2170-2180.

[6] Xiong, Jinjun, Vladimir Zolotov, Natesan Venkateswaran, and Chandu Visweswariah. "Criticality computation in parameterized statistical timing." In *Proceedings of the 43rd annual Design Automation Conference*, pp. 63-68. ACM, 2006.

[7] Zhang, Lizheng, Yuhen Hu, and C. Chung-Ping Chen. "Statistical timing analysis with path reconvergence and spatial correlations." In *Design, Automation and Test in Europe, 2006. DATE'06. Proceedings*, vol. 1, pp. 5-pp. IEEE, 2006.

[8] Zhang, Lizheng, Weijen Chen, Yuhen Hu, and Charlie Chung-ping Chen. "Statistical static timing analysis with conditional linear MAX/MIN approximation and extended canonical timing model." *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on* 25, no. 6 (2006): 1183-1191.

[9] Mogal, Hushrav D., Haifeng Qian, Sachin S. Sapatnekar, and Kia Bazargan. "Clustering based pruning for statistical criticality computation under process variations." In *Computer-Aided Design, 2007. ICCAD 2007. IEEE/ACM International Conference on*, pp. 340-343. IEEE, 2007.

[10] Shibuya, V. Izumi Nitta V. Toshiyuki, and V. Katsumi Homma. "Statistical static timing analysis technology." *Fujitsu Sci. Tech. J* 43, no. 4 (2007): 516-523.

[11] Xiong, Jinjun, Vladimir Zolotov, and Chandu Visweswariah. "Incremental criticality and yield gradients." In *Proceedings of the conference on Design, automation and test in Europe*, pp. 1130-1135. ACM, 2008.

[12] Kim, Jin Wook, Wook Kim, Park, Hyoun Soo and Young Hwan Kim. "Incremental statistical static timing analysis with gate timing yield emphasis." In *Circuits and Systems, 2008. APCCAS 2008. IEEE Asia Pacific Conference on*, pp. 1016-1019. IEEE, 2008.

[13] Blaauw, David, Kaviraj Chopra, Ashish Srivastava, and Louis Scheffer. "Statistical timing analysis: From basic principles to state of the art." *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on* 27, no. 4 (2008): 589-607.

[14] Synopsys,"Primetime Advanced OCV Technology" white paper, April 2009

[15] Xiong, Jinjun, Vladimir Zolotov, and Chandu Visweswariah. "Efficient modeling of spatial correlations for parameterized statistical static timing analysis." In *ASIC, 2009. ASICON'09. IEEE 8th International Conference on*, pp. 722-725. IEEE, 2009.

[16] Ramalingam, Anand, Ashish Kumar Singh, Sani R. Nassif, Gi-Joon Nam, Michael Orshansky, and David Z. Pan. "An accurate sparse-matrix based framework for statistical static timing analysis." *Integration, the VLSI Journal* 45, no. 4 (2012): 365-375.

[17] Chung, Jaeyong, Jinjun Xiong, Vladimir Zolotov, and Jacob A. Abraham. "Path criticality computation in parameterized statistical timing analysis using a novel operator." *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on* 31, no. 4 (2012): 497-508.

[18] Chung, Jaeyong, and Jacob A. Abraham. "Refactoring of Timing Graphs and Its Use in Capturing Topological Correlation in SSTA." *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on* 31, no. 4 (2012): 485-496.

[19] Chung, Jaeyong, and Jacob A. Abraham. "On Computing Criticality in Refactored Timing Graphs." *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on* 31, no. 12 (2012): 1935-1939.