Keynote Paper

Statistical Timing Analysis: From Basic Principles to State of the Art

David Blaauw, *Senior Member, IEEE*, Kaviraj Chopra, *Student Member, IEEE*,
Ashish Srivastava, and Lou Scheffer, *Senior Member, IEEE*

***Abstract*—Static-timing analysis (STA) has been one of the most pervasive and successful analysis engines in the design of digital circuits for the last 20 years. However, in recent years, the increased loss of predictability in semiconductor devices has raised concern over the ability of STA to effectively model statistical variations. This has resulted in extensive research in the so-called statistical STA (SSTA), which marks a significant departure from the traditional STA framework. In this paper, we review the recent developments in SSTA. We first discuss its underlying models and assumptions, then survey the major approaches, and close by discussing its remaining key challenges.**

***Index Terms*—Algorithm, circuit, performance, process variations, timing analysis.**

I. Introduction

SINCE the early 1990s, static-timing analysis (STA) has been a widely adopted tool for all facets of very-large-scale-integration chip design. STA is not only the universal timing sign-off tool but also lies at the heart of numerous timing optimization tools. The main advantage of STA over vector-based timing simulation is that it does not rely on input vectors, which can be difficult to construct and can easily miss an obscure performance-limiting path in the circuit. The widespread use of STA can be attributed to several factors: 1) The basic STA algorithm is linear in runtime with circuit size, allowing analysis of designs in excess of 10 million instances;[1] 2) the basic STA analysis is conservative in the sense that it will overestimate the delay of long paths in the circuit and underestimate the delay of short paths in the circuit. This makes the analysis "safe," guaranteeing that the design will function at least as fast as predicted and will not suffer from hold-time violations; 3) the STA algorithms have become fairly mature, addressing critical timing issues such as interconnect analysis, accurate delay modeling, false or multicycle paths, etc; and 4) delay characterization for cell libraries is clearly defined, forms an effective interface between the foundry and the design team, and is readily available.

Traditional STA tools are deterministic and compute the circuit delay for a specific process condition. Hence, all parameters that impact the delay of a circuit, such as device gate length and oxide thickness, as well as operating voltage and temperature, are assumed to be fixed and are uniformly applied to all the devices in the design. In this paper, we refer to traditional deterministic STA as DSTA. In DSTA, process variation is modeled by running the analysis multiple times, each at a different process condition. For each process condition a so-called corner file is created that specifies the delay of the gates at that process condition. By analyzing a sufficient number of process conditions the delay of the circuit under process variation can be bounded.

The fundamental weakness of DSTA is that while global shifts in the process (referred to as die-to-die variations) can be approximated by creating multiple corner files, there is no statistically rigorous method for modeling variations across a die (referred to as within-die variations).[2] However, with process scaling progressing well into the nanometer regime, process variations have become significantly more pronounced and within-die variations have become a non-negligible component of the total variation. We will show later in this paper that the inability of DSTA to model within-die variation can result in either an over- or underestimate of the circuit delay, depending on the circuit topology. Hence, DSTA's desirable property of being conservative may no longer hold for certain circuit topologies while, at the same time, DSTA may be overly pessimistic for other circuit topologies. The accuracy of DSTA in advanced processes is therefore a serious concern.

In addition to the growing importance of within-die process variations, the total number of process parameters that exhibit significant variation has also increased [1]. Hence, even the

Manuscript received September 13, 2006; revised February 12, 2007. This paper was recommended by Associate Editor S. Vrudhula.

D. Blaauw and K. Chopra are with the University of Michigan, Ann Arbor, MI 48109 USA.

A. Srivastava is with the Magma Design Automation, Austin, TX 78759 USA.

L. Scheffer is with the Cadence Design System, San Jose, CA 95134 USA.

Digital Object Identifier 10.1109/TCAD.2007.907047

[1] As discussed in Section III, the propagation of arrival times through the combinational portion of a circuit using the critical-path-method (CPM) algorithm has a runtime that is linear with circuit size. However, industrial STA tools often include methods for common-path removal in the clocking network and for false-path elimination. These methods have a higher runtime complexity than the simple CPM algorithm.

[2] While the deterministic model of gate delay as used in DSTA excludes a statistical treatment of across-die variation, industry tools have over time developed a number of methods to approximate the impact of such variations. A common method is to use a predetermined delay scaling factor for all circuit elements (delay is increased for long-path analysis and is decreased for short-path analysis). However, if the scaling factor is set to the worst-case within-die variation, the analysis becomes exceedingly pessimistic. On the other hand, lesser values cannot be proved to be conservative, negating one of the major advantages of DSTA.

0278-0070/$25.00 © 2008 IEEE

modeling of only die-to-die variations in DSTA now requires an untenable number of corner files. For instance, in addition to device parameters, interconnect parameters must be considered, and which combination of interconnect and device parameters results in the worst-case (or best-case) delay often depends on the circuit structure. In an attempt to capture the worst-case die-to-die variation for all cases, the number of corner files used in industry has risen sharply. It is now common to use more than a dozen corner files [2], whereas the number can even exceed 100, thereby increasing the effective runtime of DSTA by one order of magnitude or more.

The need for an effective modeling of process variations in timing analysis has led to extensive research in statistical STA. Some of the initial research works date back to the very introduction of timing analysis in the 1960s [3] as well as the early 1990s [4], [5]. However, the vast majority of research works on SSTA date from the last five years, with well over a hundred papers published in this research field since 2001. In this paper, we give a brief review of the different issues and approaches to SSTA. In Section II, we examine the different sources of uncertainty and their impact on circuit performance. In Section III, we present the formulation of the SSTA problem and discuss its key challenges and approaches. In Section IV, we discuss the so-called "block-based" approaches in more detail and present their strengths and weaknesses. Section V discusses the remaining key issues that must be addressed to bring SSTA to a level of maturity that approaches that of the DSTA today. We conclude this review paper in Section VI.

II. Sources of Timing Variation

In this section, we discuss the key sources of variation in timing prediction, that make timing analysis a challenging task for nanoscale digital circuits. We first discuss different types of uncertainties that arise as a design moves from specification to implementation and final operation in the field. We then focus on process variations in more detail and discuss the distinction between die-to-die and within-die variations and the source of so-called spatial correlations. Finally, we discuss the impact of different types of process variations on the timing of a circuit.

A. Process, Environmental, and Model Uncertainties

The uncertainty in the timing estimate of a design can be classified into three main categories:

1) modeling and analysis errors—inaccuracy in device models, in extraction and reduction of interconnect parasitics, and in timing-analysis algorithms;
2) manufacturing variations—uncertainty in the parameters of fabricated devices and interconnects from die to die and within a particular die;
3) operating context variations—uncertainty in the operating environment of a particular device during its lifetime, such as temperature, supply voltage, mode of operation, and lifetime wear-out.

To illustrate each of these uncertainties, consider the stages of design, from initial specification to final operation, as shown in Fig. 1. The design process starts with a broad specification of the design and then goes through several implementation steps, such as logic synthesis, buffer insertion, and place and route. At each step, timing analysis is used to guide the design process. However, timing analysis is subject to a host of inaccuracies, such as undetected false paths, cell-delay error, error in interconnect parasitics, SPICE models, etc. These modeling and analysis errors result in a deviation between the expected performance of the design and its actual performance characteristics. For instance, the STA tool might utilize a conservative delay-noise algorithm resulting in certain paths operating faster than expected.

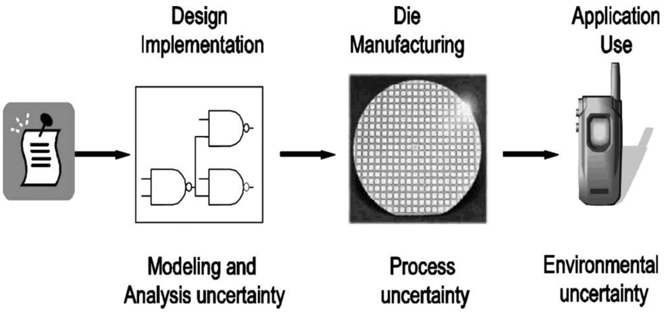

Fig. 1. Steps of the design process and their resulting timing uncertainties.

In the next stage, the design is fabricated and each individual die incurs additional manufacturing-related variations due to equipment imprecisions and process limitations. Finally, a manufactured die is used in an application such as a cell phone or a laptop. Each particular die then sees different environmental conditions, depending on its usage and location. Since environmental factors such as temperature, supply voltage, and work load affect the performance of a die, they give rise to the third class of uncertainty. To achieve the required timing specification for all used die throughout their entire lifetime, the designer must consider all three sources of uncertainty. However, a key difference between the three classes of uncertainty is that each has a sample space that lies along a different dimension. Hence, each class of uncertainty calls for a different analysis approach.

First, we recall that the sample space of an experiment or a random trial is the set of all possible outcomes. The timing uncertainty caused by modeling and analysis errors has as its sample space the set of design implementations resulting from multiple design attempts. Each design attempt results in an implementation which triggers particular inaccuracies in the models and tools, resulting in a timing distribution across this sample space. However, a design is typically implemented only once and there needs to be a high level of confidence that the constraints will be met in the first attempt. Hence, the designer is interested in the worst-case timing across this sample space. Thus, margins are typically added to the models to create sufficient confidence that they are conservative and will result in a successful implementation. Although a statistical analysis of model and analysis uncertainty is uncommon, it could aid in a more accurate computation of the delay with a specified confidence level.

In the case of process variations, the sample space is the set of manufactured die. In this case, a small portion of the sample space is allowed to fail the timing requirements since those die can be discarded after manufacturing. This considerably relaxes the timing constraints on the design and allows designers to significantly improve other performance metrics, such as power dissipation. In microprocessor design, it is common to perform so-called binning where die are targeted to different applications based on their performance level. This lessens the requirement that all or a very high percentage of the sample space meets the fastest timing constraint. Instead, each performance level in the sample space represents a different profit margin, and the total profit must be maximized.

The sample space of environmental uncertainty is across the operational life of a part and includes variations in temperature, modes of operation, executed instructions, supply voltage, lifetime wear-out, etc. Similar to model and analysis uncertainty, the chip is expected to function properly throughout its operational lifetime in all specified operating environments. Even if a design fails only under a highly unusual environmental condition, the percentage of parts that will fail at some point during their operational life can still be very high. Therefore, a pessimistic analysis is required to ensure a high confidence of correct operation throughout the entire lifetime of the part. Naturally, this approach results in a design that operates faster than necessary for much of its operational life, leading to a loss in efficiency. For instance, when a part is operating at a typical ambient temperature the device sizing or supply voltage could be relaxed, reducing power consumption. One approach to address this inefficiency is to use runtime adaptivity of the design [6], [7].

Since each of the three discussed variabilities represents orthogonal sample spaces, it is difficult to perform a combined analysis in a meaningful manner. Environmental uncertainty and uncertainty due to modeling and analysis errors are typically modeled using worst-case margins, whereas uncertainty in process is generally treated statistically. Hence, most SSTA research works, as well as this paper, focus only on modeling process variations. However, the accuracy gained by moving from DSTA to SSTA methods must be considered in light of the errors that continue to exist due to the other sources of timing error, such as analysis and modeling error, uncertainty in operating conditions, and lifetime wear-out phenomena. We discuss in the next section the sources of process variation in more detail.

B. Sources of Process Variation

1) Physical Parameters, Electrical Parameters, and Delay Variation: The semiconductor manufacturing process has become more complex, at the same time process control precision is struggling to maintain relative accuracy with continued process scaling. As a result, a number of steps throughout the manufacturing process are prone to fluctuations. These include effects due to chemical mechanical polishing (CMP), which is used to planarize insulating oxides and metal lines, optical proximity effects, which are a consequence of patterning features smaller than the wavelength of light [8]–[10], and lens imperfections in the optical system. These, as well as other numerous effects, cause variation of device and interconnect physical parameters such as gate length (or critical dimension—CD), gate-oxide thickness, channel doping concentration, interconnect thickness and height, etc., as shown in Fig. 2. Among these, CD variation and channel doping fluctuations have typically been considered as dominant factors. However, many SSTA methods model a much wider range of physical parameters. Variations in these physical parameters, in turn, result in variations in electrical device characteristics, such as the threshold voltage, the drive strength of transistors, and the resistance and capacitance of interconnects. Finally, the variations in electrical characteristics of circuit components result in delay variations of the circuit.

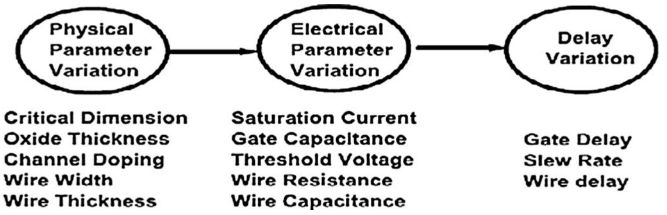

Fig. 2. Physical parameter variations resulting in electrical parameter variations, which, in turn, result in circuit-delay variations.

It is important to note that more than one electrical parameter may have a dependence on a particular physical parameter. For example, both resistance and capacitance of an interconnect are affected by variation in wire width. An increase in interconnect width reduces the separation between wires, resulting in an increased coupling capacitance while decreasing the resistance of the wire. Similarly, perturbations in the gate-oxide thickness influence the drive current, the threshold voltage, and the gate capacitance of the transistors. Dependence of two or more electrical parameters on a common physical parameter gives rise to correlation of these electrical parameters and ignoring this correlation can result in inaccurate results. For instance, if we ignore the negative correlation between capacitance and resistance, there is a nonzero probability that both resistance and capacitance are at their worst-case values. However, this is physically impossible and leads to unrealistic RC delay estimates. In [11], the authors present a method to determine the process-parameter values that result in a more realistic worst-case delay estimate.

Along similar lines, a particular equipment variation can impact multiple physical-parameter values, resulting in a correlation of the physical parameters themselves. For instance, consider the physical-parameter variations due to lens aberration. If multiple masks are illuminated with the same lens, the variation of all metal layers and even polysilicon will be correlated.[3] In Section IV, we will discuss methods for modeling correlated parameters using a smaller number of independent parameters, such as principal component analysis.

It would be ideal to model each process step in the manufacturing process to determine the variations and correlations in the physical parameters. However, such an analysis is complex and impractical due to the number of equipment-related parameters in each fabrication step and the total number of steps. Hence, most SSTA approaches have taken the physical

[3] Multiple scanners may be used to manufacture a particular part. This can reduce the discussed correlation here but may not eliminate it.

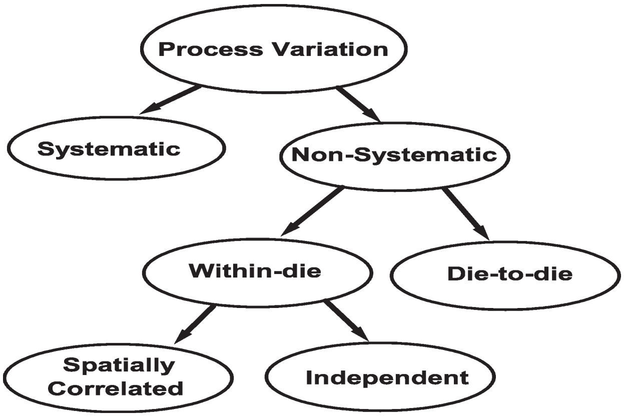

Fig. 3. Taxonomy of process variations.

parameters themselves (such as CD, doping concentration, and oxide thickness) to be the basic random variables (RVs). These variables are either assumed to be independent or to have well-understood correlations.

2) Classification of Physical-Parameter Variation: Physical-parameter variations can be classified based on whether they are deterministic or statistical and based on the spatial scale over which they operate, as shown in Fig. 3.

1) Systematic variations are components of physical-parameter variation that follow a well-understood behavior and can be predicted upfront by analyzing the designed layout. Systematic variations arise in large part from optimal proximity effects, CMP, and its associated metal fill. These layout-dependent variations can be modeled premanufacturing by performing a detailed analysis of the layout. Therefore, the impact of such variations can be accounted for using deterministic analysis at later stages of the design process [12], [13] and particularly at timing sign-off. However, since we do not have layout information early in the design process, it is common to treat these variations statistically. In addition, the models required for analysis of these systematic variations are often not available to a designer, which makes it advantageous to treat them statistically, particularly when it is unlikely that all effects will assume their worst-case values.
2) Nonsystematic or random variations represent the truly uncertain component of physical-parameter variations. They result from processes that are orthogonal to the design implementation. For these parameters, only the statistical characteristics are known at design time, and hence, they must be modeled using RVs throughout the design process. Line-edge roughness (LER) and random-dopant fluctuations (RDF) are examples of nonsystematic random sources of variation.

It is common that earlier in the design flow, both systematic and nonsystematic variations are modeled statistically. As we move through the design process and more detailed information is obtained, the systematic components can be modeled deterministically, if sufficient analysis capabilities are in place, thereby reducing the overall variability of the design.

3) Spatial Reach of Variations: Nonsystematic variations can be further analyzed by observing that different sources of variations act on different spatial scales. Some parameters shift when the equipment is loaded with a new wafer or between processing one lot of wafers to the next—this can be due to small unavoidable changes in the alignment of the wafers in the equipment, changes in the calibration of the equipment between wafer lot processing, etc. On the other hand, some shift can occur between the exposure of different reticles on a wafer, resulting in reticle-to-reticle variations. A reticle is the area of a wafer that is simultaneously exposed to the mask pattern by a scanner. The reticle is approximately 20 mm × 30 mm and will typically contain multiple copies of the same chip layout or multiple different chip layouts. At each exposure, the scanner is aligned to the previously completed process steps, giving rise to a variation in the physical parameters from one reticle to the next. Finally, some shift can occur during the reticle exposure itself. For instance, a shift in a parameter, such as laser intensity, may occur while a particular reticle is scanned leading to within-reticle variations. Another example is non-uniform etch concentration across the reticle, leading to the variation in the CD.

These different spatial scales of variation give rise to a classification of nonsystematic variations into two categories.

1) Die-to-die variations (also referred to as global or interdie variations) affect all the devices on the same die in the same way. For instance, they cause the CD of all devices on the same chip to be larger or smaller than nominal. We can see that die-to-die variations are the result of shifts in the process that occur from lot to lot, wafer to wafer, reticle to reticle, and across a reticle if the reticle contains more than one copy of a chip layout.
2) Within-die variations (also referred to as local or intradie variations) affect each device on the same die differently. In other words, some devices on a die have a smaller CD, whereas other devices on the same die have a larger CD than nominal. Within-die variations are only caused by across-reticle variations within the confines of a single chip layout.

Finally, within-die variations can be categorized into spatially correlated and independent variations as discussed as follows.

1) Spatially correlated variations. Many of the underlying processes that give rise to within-die variation change gradually from one location to the next. Hence, these processes tend to affect closely spaced devices in a similar manner, making them more likely to have similar characteristics than those placed far apart. The component of variation that exhibits such spatial dependence is known as spatially correlated variation. We discuss the modeling of spatial correlated device parameters in more detail in Section IV-B1.
2) Independent variations. The residual variability of a device that is statistically independent from all other devices and does not exhibit spatially dependent correlations is referred to as independent variation.[4] These variations include effects such as RDF and LER. It has been observed

[4] In the SSTA literature, this independent component of nonsystematic process variation is often referred to as the random component. However, this is an unfortunate misnomer since all nonsystematic variations are random.

that with continued process scaling, the contribution of independent within-die variation is increasing. Models such as those of Pelgrom *et al.* [14], which express the amount of independent variation as a function of nominal device parameters, are gaining increased importance.

C. Impact of Correlation on Circuit Delay

As discussed in the previous section, nonsystematic process variations must be modeled using RVs. Furthermore, the RVs associated with different gates in a design will be partially correlated due to the joint contributions from die-to-die, spatially correlated, and independent process-variation components. As we shall discuss in Section IV, this partial correlation creates significant difficulties for SSTA. The analysis can be substantially simplified if the RVs are assumed to be either fully correlated with a correlation coefficient of 1 or are assumed completely independent. If the RVs are assumed to be fully correlated, the variation has the same characteristics as die-to-die variation, and DSTA can be used to bound the circuit delay using a set of corner files. On the other hand, while the assumption of independence requires a statistical-analysis approach it significantly simplifies the required operations.

In this section, we investigate the error that is introduced in the timing analysis of a combinational-circuit block under either the fully correlated or independent assumption. This is useful since traditional DSTA approaches have often made the fully correlated assumption, whereas early SSTA work has made an independence assumption. We will show that, depending on the circuit topology, either assumption can yield conservative or optimistic timing estimates. In the succeeding discussion, we first consider the simple case of a single path and then treat the maximum delay of multiple paths. Finally, we note some of the complexities involved when clocking is considered.

1) Delay of a Single Path: If the gate delays along a path are independent, then they tend to average out in the overall path delay. For example, let a path have n gates connected in series with each gate having an independent normal-delay distribution $P_1, P_2, \ldots, P_n$ with the same mean μ and standard deviation σ. The σ/μ ratio of the path delay relative to that of the gate delay is given by

$$\left(\frac{\sigma}{\mu}\right)_{\text{path}} = \frac{1}{\sqrt{n}}\left(\frac{\sigma}{\mu}\right)_{\text{gate}}. \tag{1}$$

However, if the gate-delay distributions P_i are correlated with a correlation coefficient ρ, the σ/μ ratio for the path delay becomes

$$\left(\frac{\sigma}{\mu}\right)_{\text{path}} = \sqrt{\frac{1+\rho(n-1)}{n}}\left(\frac{\sigma}{\mu}\right)_{\text{gate}}. \tag{2}$$

In both cases, the standard deviation of the path delay increases with the number of gates in the path. However, the ratio σ/μ scales as $1/\sqrt{(n)}$ with the assumption of independence, whereas it remains constant under the fully correlated assumption ($\rho = 1$). Hence, an assumption that the partially correlated RVs along a path are fully correlated will overestimate the spread of path delay. The delay specified at a confidence point greater than 50%[5] will be overestimated, resulting in a pessimistic analysis. On the other hand, assuming all variations to be independent along a path results in an optimistic estimate of delay (at yield points greater than 50%).

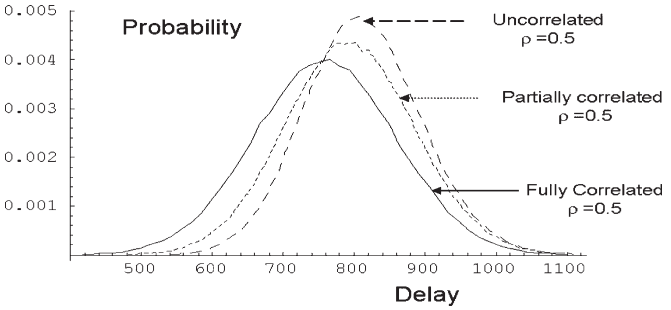

Fig. 4. Maximum of two normal-delay distributions with identical mean and variance having different correlations.

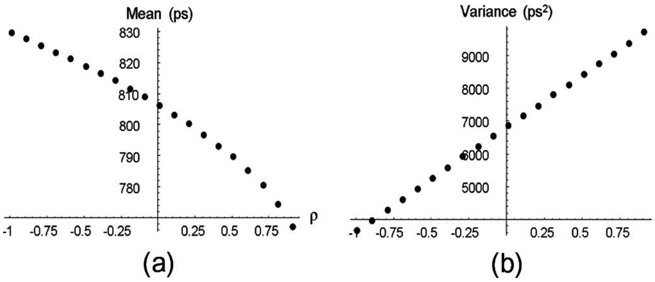

Fig. 5. Mean and variance of the maximum of two identical normal RVs as a function of correlation coefficients.

2) Maximum of Multiple Paths: The delay of a combinational-circuit block is obtained by taking the maximum of the delays of all the paths in the circuit. This maximum again depends on the correlation of the gate delays. To illustrate this, we consider a circuit consisting of two paths, each with normal distributed delays with a mean of 750 ps and a standard deviation of 100 ps. If we assume that all of the gate delays in the circuit are fully correlated with $\rho = 1$, the two path delays will also be perfectly correlated. On the other hand, if the gate delays are assumed to be independent, the two path delays will be independent. Fig. 4 shows the probability distribution of the maximum of the two path delays assuming perfect correlation ($\rho = 1$), partial correlation ($\rho = 0.5$), and independent path delays ($\rho = 0$). It can be seen that the probability distribution for the independent case always gives a higher delay than that in the two positively correlated cases. Hence, the independence assumption will overestimate delay and vice versa. This concept is mathematically known as Slepian's inequality. Intuitively, when the two path delays are independent, the number of cases in the sample space, where at least one of the two delays is toward the high end of the distribution, is much greater compared to the correlated case.

Fig. 5 shows the mean and variance of the maximum of the two path delays as a function of their correlation coefficient. The mean of the maximum delay decreases with increased

[5] For long-path delay analysis, delay is typically computed at confidence points greater than 50%, whereas, for short-path delay analysis, the confidence point is typically placed $< 50\%$. In both cases, overestimation of ρ results in a pessimistic analysis.

correlation, whereas the standard deviation increases. Therefore, the independence assumption will underestimate the delay spread. Note also that the maximum of the two identical normal path-delay distributions does not result in a normal distribution except under the fully correlated assumption. We will discuss this effect in more detail in Section III.

Given a combinational-circuit block, the overall circuit delay maybe either over- or underestimated by both the fully correlated and independence assumptions. The fully correlated assumption will overestimate the delay of individual paths, whereas it will underestimate the maximum of those path delays. Hence, the final outcome depends on the topology of the circuit. If the logic depth of the circuit is large while, at the same time, only a few paths are critical, the overestimation along the critical paths will dominate, resulting in a pessimistic analysis result. On the other hand, for circuits with shallow logic depth and highly balanced paths, the underestimation occurring in the maximum computation will dominate, and the analysis will be optimistic under the fully correlated assumption. The inverse is true for the independence assumption.

Finally, when considering sequential circuits, the delay variation in the buffered clock tree must be considered. In general, the fully correlated assumption will underestimate the variation in the arrival times at the leaf nodes of the clock tree, which will tend to overestimate circuit performance. However, we must also consider the correlation between the delays in the combinational logic and the clock tree, in which case the analysis becomes more complex.

III. Problem Formulation and Basic Solution Approaches

A. Problem Formulation

The traditional DSTA procedure abstracts a timing graph from a combinational circuit. The nodes of the timing graph represent primary inputs/outputs of the circuit and gate input/output pins. Its edges represent the timing elements of the circuit, namely, the gate input-pin–output-pin delay and wire delay from a driver to a receiver, as shown in Fig. 6. The weight on these edges represents the delay of the corresponding timing element. For a combinational circuit, it is convenient to connect all primary inputs to a virtual source node with virtual edges having weight equal to the input arrival times. Similarly, all the primary outputs are connected to a virtual sink node through virtual edges with weights representing the required arrival times. The resulting timing graph, therefore, has a single source and sink node.

A similar timing graph can be constructed for sequential circuits. Fig. 7 shows the additional timing elements pertaining to a clock network (i.e., the launch and capture paths of the clock tree) and the sequential elements. In the corresponding timing graph, the virtual source node corresponds to the input driver of the on-chip clock network. The clock-driver delays and interconnect delays on the launch path, the clock-to-q delay, and the setup times of the sequential elements are again modeled using weights on their corresponding graph edges. Similarly, the virtual sink node also corresponds to the clock input driver, and the capture path is represented with nodes and edges in the timing graph. In this case, however, the weights of edges corresponding to the capture path are assigned with negative delay values as opposed to the positive values for the launch path. Apart from this distinction, the timing graphs for flip-flop-based sequential circuits are a direct extension of those for the combinational circuits and can be analyzed with the same timing algorithms. However, significant complications arise when transparent latches are used in place of flip-flops or when the launch and capture paths of the clock tree share the same drivers, as is common.

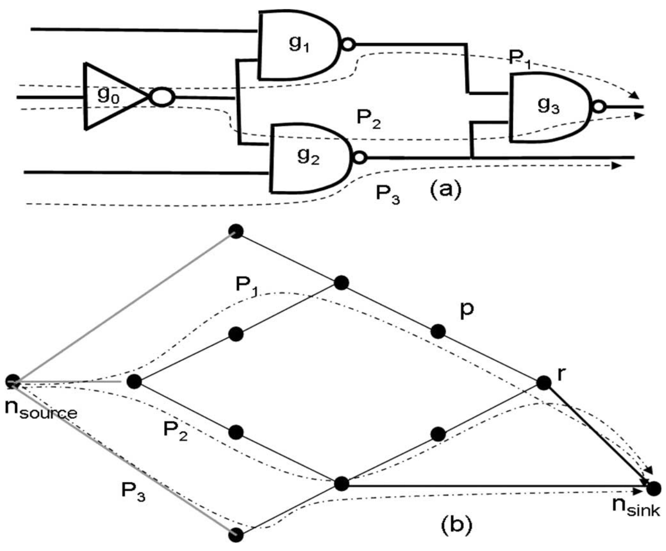

Fig. 6. Example circuit in (a) and its timing graph in (b).

As discussed in Section II, device parameters such as gate length, doping concentration, and metal thickness must be treated as RVs due to process variation. The delay of each edge, being a function of these parameters, also becomes an RV. This allows us to extend the concept of the traditional timing graph to a statistical timing graph defined as follows.

Definition: A timing graph $G = \{N, E, n_s, n_f\}$ is a directed graph having exactly one source node n_s and one sink node n_f, where N is a set of nodes, and E is a set of edges. The weight associated with an edge corresponds to either the gate delay or the interconnect delay. The timing graph is said to be a statistical timing graph if ith edge weight d_i is an RV.

The arrival time at the source node of the timing graph typically has a deterministic zero value. This reflects the fact that in combinational timing graphs, clock-tree skew is not represented, whereas in sequential circuits, the source node is pulled back to a common point on the launching and capturing clock paths.[6] In traditional DSTA, the most basic goal of the analysis is to find the maximum delay between the source node and the sink node of a timing graph, which is the delay of the longest path in the circuit. When modeling process-induced delay variations, the sample space is the set of all manufactured dies. In this case, the device parameters will have different values across this sample space, hence the critical path and its delay will change from one die to the next. Therefore, the delay

[6] Note that a deterministic value at the source node of a sequential timing graph does not account for jitter from the Phase Locked Loop (PLL) or other sources.

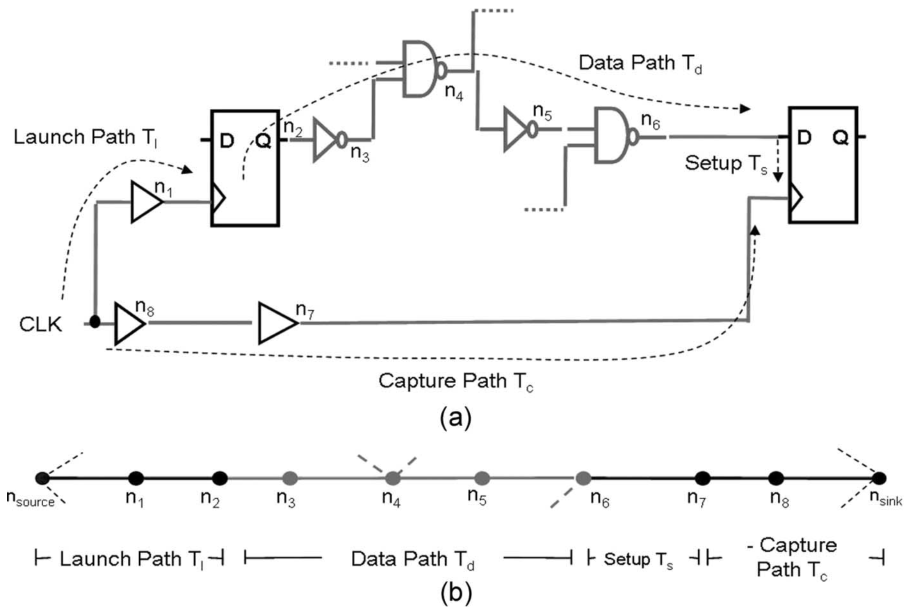

Fig. 7. Timing elements of a sequential circuit path (a) and its timing graph (b).

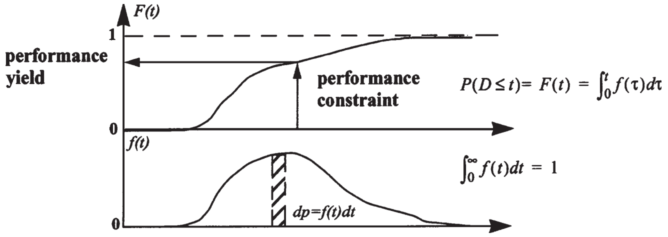

Fig. 8. PDF and CDF.

of the circuit is also an RV, and the first task of SSTA is to compute the characteristics of this RV. This is performed by computing its probability-distribution function (PDF) or cumulative-distribution function (CDF) (see Fig. 8). Alternatively, only specific statistical characteristics of the distribution, such as its mean and standard deviation, can be computed. Note that the CDF and the PDF can be derived from one another through differentiation and integration. Given the CDF of circuit delay of a design and the required performance constraint the anticipated yield can be determined from the CDF. Conversely, given the CDF of the circuit delay and the required yield, the maximum frequency at which the set of yielding chips can be operated at can be found.

Definition: Let a path p_i be a set of ordered edges from the source node to the sink node in G, and let D_i be the path-length distribution of p_i, computed as the sum of the weights d for all edges k on the path. Finding the distribution of $D_{\max} = \text{maximum}(D_1, \ldots, D_i, \ldots, D_{n\text{paths}})$ among all paths (indexed from 1 to n paths) in the graph G is referred to as the SSTA problem of a circuit.

Similar to traditional DSTA, we can formulate the SSTA problem as that of finding the latest arrival-time distribution at the sink node in the timing graph [15], [16]. The latest arrival-time distribution at the sink node can be found by propagating the arrival time from the source node through the timing edges while computing the latest arrival-time at every node in topological order. Subsequently, the latest arrival-time distribution at the sink node is the circuit-delay distribution. It is worth noting that the basic DSTA algorithm is based on the project-planning technique known as the CPM and involves a simple topological traversal [17]. Likewise, the basic SSTA formulation for circuit designs was first motivated from the project evaluation and review technique (PERT) literature [3], [18]. However, in contrast to DSTA, PERT was shown to be an N-P complete problem [19].

In addition to the problem of finding the delay of the circuit, which we have posed as the basic SSTA problem, it is also key to improve this delay when the timing requirements are not met. Hence, DSTA methods typically report the slack at each node in the circuit, in addition to the circuit delay and critical paths. The slack of a node is the difference between the latest time a signal can arrive at that node, such that the timing constraints of the circuit are satisfied (referred to as the required time), and the actual latest arrival time of the signal at that node [20]. Similar to the circuit delay, the slack of a node is an RV in the SSTA formulation. Due to space limitation, we will not discuss efficient methods for slack computation in this paper but refer to the pertinent literature in Section V. We also will not discuss latch-based sequential timing analysis, which involves multiple-phase clocks, cycle stealing, clock-schedule verification, etc. Methods for statistical sequential timing analysis using latches and clock-skew analysis can be found in [21]–[25].

B. Challenges in SSTA

The statistical formulation of timing analysis introduces several new modeling and algorithmic issues that make SSTA a complex and enduring research topic [26]. In this section, we introduce some of these issues, as well as the relevant SSTA terminology.

1) Topological Correlation: Paths that start with one or more common edges after which the paths separate and join again at a later node are called reconvergent paths and the node at which these paths reconverge is called the reconvergent node. For instance, in Fig. 6, the two paths P_1 and P_2 share the same first edge (corresponding to gate g_1) and reconverge at the output of gate g_0 (node r). In such case, the input arrival times at the reconvergent node become dependent on each other because of the shared edge delay. This dependence leads to so-called topological correlation between the arrival times and complicates the maximum operation at the reconvergent node. To perform accurate analysis, the SSTA algorithm must capture and propagate this correlation so that it is correctly accounted for during the computation of the maximum function.

2) Spatial Correlation: As discussed in Section II-B, within-die variation of the physical device parameters often exhibits spatial correlation, giving rise to correlation between the gate delays. Hence, if the gates that comprise two paths have spatially correlated device parameters they will have correlated path delays. In this way, correlation can be introduced between paths that do not share any common timing edges. For instance, in Fig. 6, the paths P_1 and P_3 do not share any common delay edges, but if gates g_1 and g_2 are within close proximity on the die, their spatially correlated delays can give rise to correlation between the two path delays. Hence, spatial correlation of the arrival times must be captured and propagated during SSTA so that it is correctly accounted for during the maximum operation. Spatial correlation also impacts the sum operation. For example, if in Fig. 6, gates g_1 and g_3 have spatially correlated delays then the arrival time at node p will be correlated with the delay of gate g_3.

While topological correlation only affects the maximum operation, spatial correlation affects both the sum operation and the maximum operation. This raises two fundamental challenges for SSTA: 1) how to model gate delays and arrival times such that the spatial correlation of the underlying device parameters can be expressed; and 2) given a model of the spatial correlation, how to propagate and preserve the correlation information while performing the sum and maximum operations in SSTA. A common approach to this problem has been to represent delay in terms of the device-parameter-space basis, which is common to all gate delays. This approach is discussed in more detail in Section IV.

3) Nonnormal Process Parameters and Nonlinear Delay Models: Normal or Gaussian distributions are found to be the most commonly observed distributions for RVs, and a number of elegant analytical results exist for them in the statistics literature. Hence, most of the initial work in SSTA assumed normal distributions for physical device parameters, electrical device parameters, gate delays, and arrival times. However, some physical device parameters may have significantly nonnormal distributions. In this section, we discuss the source and impact of such nonnormal distributions.

An example of a nonnormal device parameter is CD (or gate length) due to the variation in depth of focus (DOF). As a result of equipment limitations and nonplanarity of the die, the focus point of the exposed image on the die exhibits some amount of variation. This impacts the development of the photoresist layer and consequently impacts the CD of the device. However, both large and small values of the DOF result in an underdevelopment of the photoresist and the dependence of CD on the DOF is nonlinear. Even if the variation of DOF is normal, the CD variation will decidedly be nonnormal. As shown in Fig. 9, the PDF of CD is clearly (negatively) skewed and nonnormal.[7]

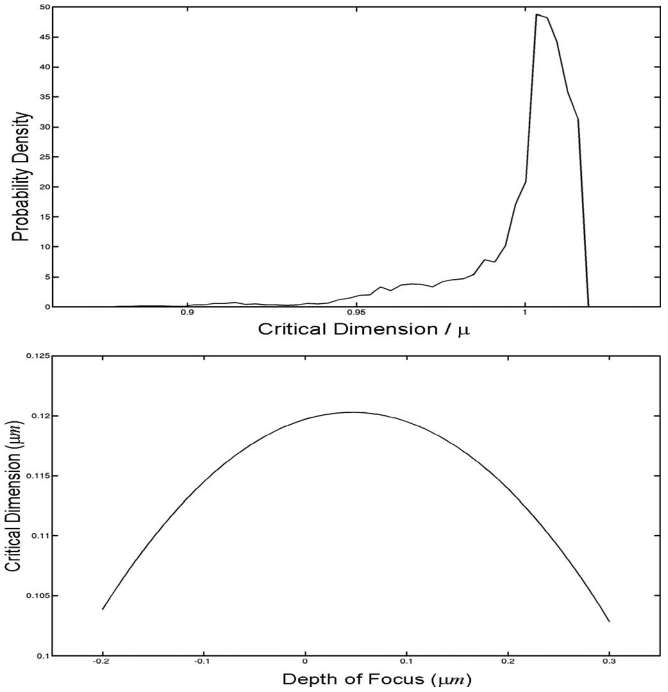

Fig. 9. Nonnormal CD distribution due to nonlinear dependence of CD on DOF.

Even if the physical device parameters are indeed normally distributed (e.g., doping concentration has a normal distribution), the dependence of the electrical device parameters and gate delay on these physical parameters may not be linear, giving rise to nonnormal gate delays. Initial work in modeling spatial correlations [27]–[29] used a first-order delay model which assumed a linear dependence of the gate delay on physical device parameters. If the variations are small, this linear approximation is justified, as the error introduced by ignoring higher order terms is negligible. However, with reduction of geometries, process variation is becoming more pronounced, and the linear approximation may not be accurate for some parameters.

Nonnormal delay and arrival-time distributions introduce significant challenges for efficient SSTA. While this is a relatively new area of research, several researchers have proposed approaches to address this issue [30]–[35]. Finally, it should be noted that apart from the difficulty of modeling the nonnormality of an individual RV, the dependence between two nonnormal RVs is no longer expressed by a simple correlation factor. This further complicates the correct treatment of topological and spatial correlations.

4) Skewness Due to Maximum Operation: Even if gate delays are assumed to be normal, SSTA has to cope with the

[7] A probability distribution is said to have negative skewness if it has a long tail in the negative direction of the RV, such as the CD distribution shown in Fig. 9. Conversely, a positive skewness indicates a long tail in the positive direction.

fact that the maximum operation is an inherently nonlinear function. The maximum of two normal arrival times will result in a nonnormal arrival time that is typically positively skewed.[8] In addition, the nonnormal arrival-time distribution produced at one node is the input to the maximum computation at downstream nodes. Hence, a maximum operation that can operate on nonnormal arrival times is required.

Most of the existing approaches ignore the skewness introduced by the maximum operation and approximate the arrival times with normal distributions. The error of this normal approximation is larger if the input arrival times have similar means and dissimilar variances [36]. In other words, the error is most pronounced when two converging paths have nominally balanced path delays, but one path has a tighter delay distribution than the other. This can occur in a circuit when two paths with equal nominal delay have a different number of gates or when the correlation among their gates differs. Another example is when one path is dominated by interconnect delay while the other is dominated by gate delay.

An example of such two delay distributions is shown in Fig. 10(a). Intuitively, we can see that RV B will dominate the maximum delay for values greater than its mean since B has significantly higher probabilities in this range. For delay values below the mean, RV A will dominate. Since A and B have different variance, skew is introduced in their maximum. For two input distributions that have identical means and variances, the resulting maximum exhibits smaller skewness [Fig. 10(b)]. Finally, Fig. 10(c) shows that if the means of the input distributions are significantly different, the resulting maximum is entirely dominated by one distribution, and skew is negligible.

The aforementioned issues address four basic challenges in SSTA, which have received significant attention in the literature. However, many other critical challenges to the development of a mature SSTA tool remain. For instance, the availability of statistical data remains difficult. This, and other challenges in SSTA, will be discussed in Section V.

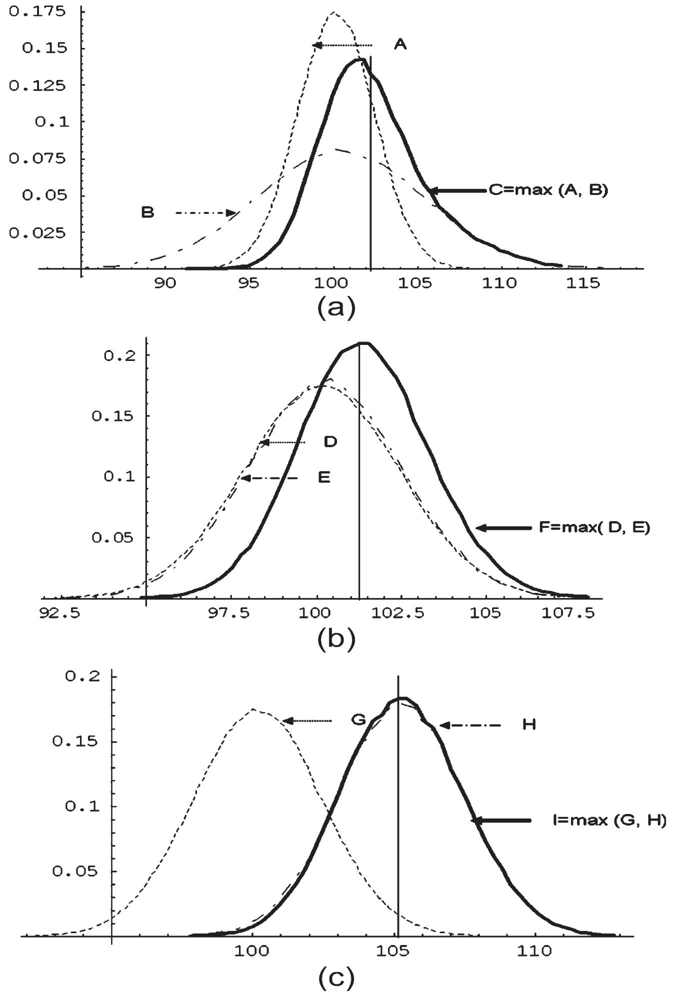

Fig. 10. Skewness due to nonlinear maximum operation for different input distributions.

C. SSTA Solution Approaches

We now give a brief overview of the principle approaches to SSTA, moving from traditional methods to more recent approaches.

1) Numerical-Integration Method: The simplest SSTA approach follows directly from the problem definition given in Section III-A. A numerical integration over the process-parameter space is performed to compute the yield of the circuit for a particular delay. Typically, the delay of a set of critical paths is expressed as a linear function of the physical device parameters and a feasible region in this parameter space is defined by the desired circuit delay. This region is then numerically integrated, exploring all possible permutations of physical device-parameter values that lie in the feasible region. Efficient numerical-integration methods were proposed in [37]. The advantage of this method is that it is completely general and process variation with any type of distribution and correlation can be modeled. However, it can be quite expensive in runtime, in particular for balanced circuits with a large number of critical paths.

2) Monte Carlo Methods: The second general approach performs a statistical sampling of the sample space using Monte Carlo simulation, based on the Metropolis sampling algorithm [38]. Instead of explicitly enumerating the entire sample space, the key idea is to identify the regions of significant probability and to sufficiently sample these regions. Using the PDF of the physical device parameters, a number of samples are drawn. For each sample, the circuit delay is computed using the traditional DSTA methods. Thereafter, by evaluating a fraction of samples that meet the timing constraint, an estimate of timing yield is found. If a sufficient number of samples are drawn, the estimation error is small. By sweeping the timing constraint and finding the yield for each value, the entire circuit-delay distribution can be found.

As with numerical integration, the Monte Carlo approach has the advantage of being completely general. Furthermore, it is based on existing mature DSTA methods and performs significantly faster than the numerical integration-based methods. However, since DSTA is in the inner loop of the Monte Carlo

[8] It is possible to obtain much more complex distributions, such as bimodal distributions, even when the input parameters remain normal. While such occurrence is rare, they introduce significant modeling difficulties.

simulation, the runtime can still be significant, particularly if a fully featured industrial DSTA tool is used. Using the Monte Carlo simulation, it is also difficult to perform incremental analysis after a designer makes a small change to the circuit. It has been shown that the performance of Monte Carlo techniques can be improved using methods such as importance sampling [2], [39]–[41]. However, more research is required to examine if fast sampling techniques can be effective for SSTA.

3) Probabilistic Analysis Methods: Both previous approaches are based on sample-space enumeration. In contrast, probabilistic methods explicitly model gate delay and arrival times with RVs. These methods typically propagate arrival times through the timing graph by performing statistical sum and maximum operations.[9] They can be classified into two broad classes: 1) path-based approaches; and 2) block-based approaches. The key difference between the two approaches is where in the algorithm the maximum function is invoked.

Path-based approaches: In path-based SSTA algorithms, a set of paths, which is likely to become critical, is identified, and a statistical analysis is performed over these paths to approximate the circuit-delay distribution. First, the delay distribution of each path is found by summing the delay of all its edges. Assuming normal gate delays, the path-delay distribution is normal and can be analytically computed [42]–[44]. The overall circuit-delay distribution is then found by performing a statistical maximum operation over all the path delays (discussed in more detail in Section IV).

The basic advantage of this approach is that the analysis is clearly split into two parts—the computation of path delays followed by the statistical maximum operation over these path delays. Hence, much of the initial research in SSTA was focused on path-based approaches [5], [42], [43], [45]–[48]. Clearly, the difficulty with the approach is how to rigorously find a subset of candidate paths such that no path that has significant probability of being critical in the parameter space is excluded. In addition, for balanced circuits, the number of paths that must be considered can be very high. Therefore, most of the later research has focused on the block-based approaches.

One of the methods that fall in the path-based category approximates the statistical delay of a circuit early in the design process when the exact gate-level implementation is not yet known [49], [50]. In this work, the circuit is modeled using a set of generic paths whose specifications are provided by the designer. The method also determines the settings of the transistor-level electrical parameters that give a specific yield goal. These settings can then be used in a traditional deterministic timing verification flow. The usefulness of applying SSTA methods early in the design process, when exact gate-level implementations are not yet available, depends on the relative magnitude of the delay uncertainty introduced by process variations versus the uncertainty due to the undetermined circuit implementation.

Block-based approaches: The block-based methods follow the DSTA algorithm more closely and traverse the circuit graph in a topological manner. The arrival time at each node is computed using two basic operations: 1) For all fan-in edges of a particular node, the edge delay is added to the arrival-time at the source node of the edge using the sum operation; and 2) given these resulting arrival times, the final arrival time at the node is computed using the maximum operation. Hence, the block-based SSTA methods propagate exactly two arrival times (a rise and a fall arrival time) at each circuit node, resulting in a runtime that is linear with circuit size. The computation of the sum function is typically not difficult; however, finding the statistical maximum of two correlated arrival times is not trivial.

Due to its runtime advantage, many current research and commercial efforts have taken the block-based approach. Furthermore, unlike other approaches, the block-based approach lends itself to incremental analysis which is advantageous for diagnostic/optimization applications. In block-based SSTA methods, the result of the maximum operation performed at one node is the input to the maximum operation which is performed at downstream nodes. It is therefore essential that the sum and maximum operations preserve the correlation information of the arrival times so that this information is available at later operations. Furthermore, the skewness introduced by the maximum operation must be considered.

IV. Block-Based SSTA

In this section we discuss block-based SSTA methods in more detail. The different methods are presented in order of increasing complexity. We start with simpler early methods that were based on a normal independent approximation of the arrival times. We then discuss methods that model topological correlation due to reconvergence of arrival times. This is followed by a number of methods that account for spatial within-die variations. Finally, we briefly survey more recently proposed nonlinear and nonnormal block-based methods.

A. Distribution Propagation Approaches (Gate-Delay Space)

Initial efforts in block-based SSTA approaches focused on directly representing gate delays with RVs characterized by their distribution or statistical characteristics. The common technique employed by all these approaches is to explicitly propagate the arrival-time distributions through the timing graph. This is achieved by employing a statistical sum operator to compute the sum of the timing arc delay and the source-node arrival-time distribution. In the case of multifan-in nodes, a statistical maximum operator is also applied to the arrival times corresponding to different fan-in edges of a node.

A basic block-based SSTA algorithm based on a PERT-like traversal was first given in [3]. Later, Berkelaar [51] presented a linear-runtime algorithm for propagating mean and variance of timing variables. In this approach, both gate delays and latest arrival-time distributions are assumed to be independent normal RVs. Based on these simplifying assumptions, the sum and maximum of arrival-time RVs are computed using analytical results for the normal RVs.

[9] The minimum operation is also needed for the computation of the shortest path, clock skew, and slack computations. However, it can be derived from the maximum operation.

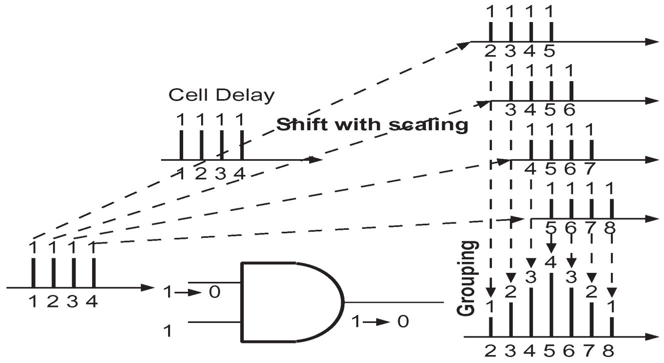

Fig. 11. Shift with scaling and grouping techniques to perform convolution of input and gate-delay PDFs to compute the output-delay PDF.

In [52], the authors extend this analytical approach to handle topological correlation due to reconvergent paths, and correlation between edge delays that correspond to the same gate, at the cost of increased complexity. The approach uses the statistical sum operation to reduce series edges in the timing graph to a single edge. At each step of the reduction, the correlation of the reduced edge with the edges with which it has nonzero correlation is recomputed. A similar reduction procedure is then performed for parallel edges using the statistical maximum operation under the normal assumption using the analytical results given in [36]. This maximum operation is explained in more detail in the following section. The proposed approach limits the number of edges whose correlation information is stored in the memory by identifying those nodes whose correlation information is no longer required. This approach was extended in [53] and [54] by assigning a level to each node in the directed acyclic graph (DAG) using a depth-first search. The level is used to identify the nodes whose correlation information can be discarded at each stage of the arrival-time propagation.

In [55]–[57], the authors propose an alternative discrete representation for relaxing the normal-distribution assumption. The gate delays are now modeled as discrete delay distributions that are generated by sampling a continuous distribution. Note that the discrete PDFs are renormalized after sampling to ensure that the sum of the probabilities for the discrete events is equal to one.

The approach then utilizes discrete sum and maximum operations for arrival-time propagation. In the case of a degenerate or deterministic input-delay distribution, the sum operation is simple, and the output-delay PDF is obtained by simply shifting the gate-delay distribution by the input delay. However, in the case where the input-delay PDF is nondegenerate, a set of shifted output-delay distributions is generated, as shown in Fig. 11. Each of these shifted PDFs corresponds to a discrete event from the input-delay PDF. This set of shifted PDFs is then combined using Bayes' theorem—the shifted PDFs are first scaled, where the scaling factor is the probability of the associated discrete input event. The scaled events are then grouped by summing the probability at each of the discrete time points. The actual probability of an event can be obtained by dividing the total value for each discrete point of the PDF by the sum of the numbers corresponding to all the events in each discrete PDF. The overall computation can be succinctly expressed as

$$f_s(t) = \sum_{i=-\infty}^{\infty} f_x(i) f_y(i-t) = f_x(t) \star f_y(t) \tag{3}$$

where $s = x + y$, and implies that the PDF of the sum of two RVs can be expressed as a convolution of their PDFs.

The statistical maximum is computed using the relation

$$f_z(t) = F_x(t)\, f_x(t) + F_y(t) f_y(t) \tag{4}$$

where $z = \text{maximum}(x, y)$, f and F represent the PDF and CDF of the RV, respectively, and x and y are assumed to be independent. The previous equation expresses mathematically that the probability that the maximum of two discrete RVs has a value t_0 is equal to the probability that one of the RVs has a value equal to t_0 and the other has a value less than or equal to t_0.

For handling topological correlation due to reconvergent paths, a partitioning-based approach is used to decompose the circuit into the so-called supergates. Each supergate is a subcircuit with a single fan-out and one or more inputs, all of which are statistically independent. The discrete events at the inputs of the supergates are propagated separately, and the resulting distributions are finally combined at the output of the supergate using Bayes' theorem. The process of separately propagating each of the discrete events of the PDFs is referred to as enumeration. Special care has to be taken in the case where a multifan-out node lies in the fan-out cone of another multifan-out node. Unfortunately, the runtime complexity of this algorithm depends on the circuit structure and is exponential in the worst case.

The authors in [15], [58], and [59] extend the work on handling topological correlation while using the same discrete framework for representing PDFs. The authors present an approach to determine the minimum set of nodes, which needs to be enumerated to handle reconvergence exactly. As expected, the worst-case computational complexity of enumeration remains exponential. Nevertheless, the authors show the useful property that ignoring topological correlation results in a distribution that is a stochastic upper bound on the exact distribution of the circuit delay. A stochastic upper bound of a delay distribution with CDF $P(t)$ is a distribution whose CDF $Q(t)$ has a value which is always smaller than or equal to $P(t)$ for all values of t, as shown in Fig. 12. Such an upper bound results in a pessimistic estimate of the timing yield of the circuit at a given performance target.

Based on this result, the authors developed a linear-runtime method for computing both lower and upper bounds on the exact delay distribution of the circuit. These bounds are then used to obtain an estimate of the circuit delay at a desired confidence point, as well as the accuracy of the bounds. In the case when the bounds are not sufficiently close to each other, a heuristic method is used to iteratively improve the bounds using selective enumeration of a subset of the nodes. The results presented in [15] showed that performing enumeration at a small set of carefully selected nodes leads to a significant improvement in the quality of the bounds. This is due to the

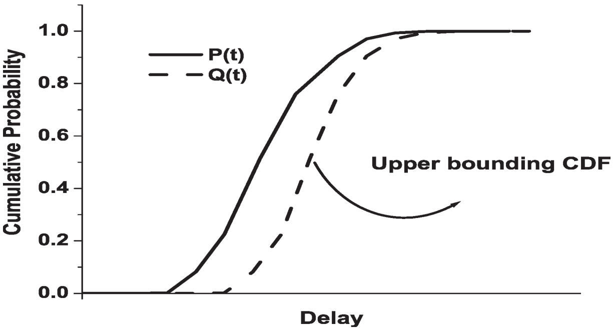

Fig. 12. Upper bound of a delay CDF provides a conservative estimate of circuit delay for a given timing yield.

fact that correlation between two arrival times only impacts the maximum computation if the two arrival times have comparable means. Hence, the correlation between two arrival times that are substantially shifted can be ignored without incurring significant error.

In a related work [60], a Bayesian-network-based approach for representing the statistical timing graph is presented for handing topological correlations. The Bayesian-network formulation prescribes an efficient method to factorize the joint distribution over the entire timing graph into an optimal set of factors. Although the worst-case runtime complexity of such an approach remains exponential, the complexity grows exponentially with the size of the largest clique in the circuit, which, in practice, is found to grow much more slowly than the circuit size.

In [61], the authors modeled arrival times as CDFs and delays as PDFs. Using a piecewise linear model for CDFs and PDFs, they present a computationally simple expression for performing the sum and maximum operations. Furthermore, they also presented a method for handling reconvergent fan-outs using a dependence list associated with every arrival time, which are propagated through the circuit and pruned for increased efficiency. Using error budgeting, an approach to optimize the runtime of this method was presented in [62]. A method to generate device-level discrete delay distributions from the underlying process-parameter distribution was presented in [63].

B. Dependence Propagation Approaches (Parameter Space)

In the previous section, we discussed techniques that consider topological correlations. The next crucial step in the development of block-based SSTA was to account for spatial correlation of the underlying physical device parameters. The basic difference between the two cases is that the correlation among arrival times now originates from the correlation of the device parameters. In addition, an arrival time at the input of a gate can be correlated with the delay of the gate itself, impacting the sum operation in addition to the maximum operation.

In the distribution propagation approaches, the gate delays are the basic RVs in the formulation. However, to model the correlation in the physical device parameters, it is necessary to model these device parameters themselves as the basic RVs. The delay of the gates is therefore expressed as a function (linear or nonlinear) of the physical or electrical device parameters. It is this functional form that expresses the dependence of the gate delays on the device parameters, which is propagated through the circuit. This concept of representing delay dependences with a parametric delay model was first introduced in [64]. To enable such techniques, it is necessary to develop a model of the spatial correlation of the device parameters. Therefore, we first discuss some of the models for expressing the correlation of device parameters and then show how these can be used to compute the final circuit delay.

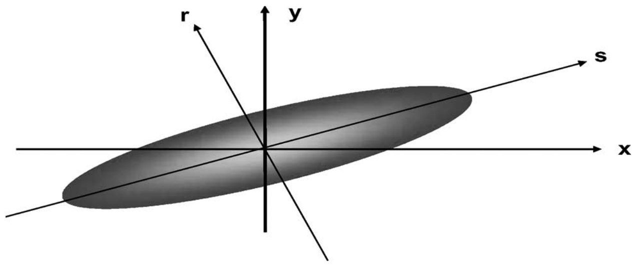

Fig. 13. Principal components of two positively correlated RVs.

1) Correlation Models: To exactly model spatial correlation between the physical parameters of two devices, a separate RV is required for each device. However, the correlation between two devices is generally a slow monotonically decreasing function of their separation, decaying over distances of hundreds of micrometers. Therefore, simplified correlation structures using a grid model [29] or quadtree model [28] have been proposed. These models allow the correlation among gates of the die to be expressed using a smaller set of RVs.

In a grid model, the overall die area is divided using a square grid. It is assumed that the grid size is chosen such that all gates within a single square on the grid can be assumed to have perfectly correlated spatial variations. Let us now consider the RVs required to model variations in a given process parameter. Each square in the grid corresponds to an RV of a device parameter, which has correlations to all other RVs corresponding to the other squares. To simplify the correlation structure of the RVs, this set of correlated RVs is mapped to another set of mutually independent RVs with zero mean and unit variance using the principal components of the original set of correlated RVs. The original RVs are then expressed as a linear combination of the principal components. These principal components can be obtained by performing an eigenvalue decomposition of the correlation matrix, as explained in more detail in [65].

Intuitively, this is shown in Fig. 13 where the distribution of two correlated jointly normal RVs A and B is shown. In the scatter plot, the x-axis is the value of A, whereas the y-axis is the value of B. If A and B were independent, the scatter plot would form a perfect circle or a horizontal or vertical oval. The diagonal distribution shown indicates positive correlation between A and B since large values of A tend to correspond to large values of B. The principal-component-analysis (PCA) method expresses A and B using two new RVs C and D, using the rotated axes r and s. RVs A and B can be expressed using a linear combination of C and D. Furthermore, the rotation of r and s ensures that C and D are independent.

It is important to note that constructing the correlation matrix directly from a distance-based correlation function may result

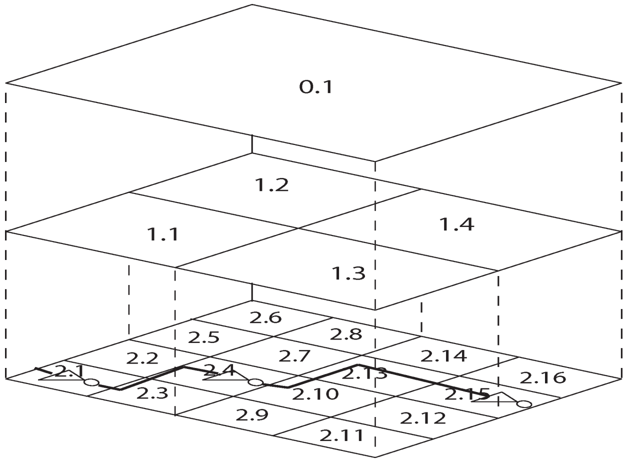

Fig. 14. Modeling spatial correlation using quadtree partitioning. The numbering of regions in different levels is shown in the figure. A region (i, j) intersects the regions $(i+1, 4j-3)$–$(i+1, 4j)$.

in a nonpositive-definite matrix. Furthermore, the correlation matrix must be positive definite and that this condition may be violated if the matrix is constructed from an arbitrary distance-based function or from measured data (especially if noisy). Hence, some techniques, such as replacing all negative eigenvalues by zero, may need to be used. This problem has been investigated in [66]–[71].

The quadtree model proposed in [28] and [43] also partitions the overall die area into a number of regions. However, instead of using PCA to express the correlated components of variations, it uses an additive approach to consider the spatial dependence of process parameters. This is achieved by recursively dividing the area of the die into four equal parts, which is known as quadtree partitioning. As the regions of the die are recursively divided into parts, the number of parts at each level increases by a factor of four, as shown in Fig. 14. Each partition, at all levels of the quadtree, is assigned with an independent RV. The spatially correlated variation associated with a gate is then defined to be the sum of the RV associated with the lowest level partition that contains the gate and the RVs at each of the higher partitioning levels that intersects the position of the gate. The correlation among gates arises from the sharing of some of the RVs at higher levels of the quadtree. The number of shared RVs depends on the distance between the two gates. Moreover, a larger fraction of the variance can be allocated at higher levels of the quadtree if the parameter is known to be more strongly correlated over large distances. In [72], a method for determining the values of the RVs associated with each partition is presented.

An alternative grid-based model was proposed in [33] where only four RVs are used to model the correlation structure. The four RVs are assumed to be associated with the four corners of a die, and the RVs associated with the gates inside the design are represented as a weighted sum of these four RVs, where the weighting coefficients are functions of the distance between the position of a gate and each of the four corners of the die.

In [73], the authors use the Karhunen–Loeve expansion (KLE) to express the distance-based correlation function in terms of a set of RVs and an orthonormal set of deterministic functions related to the position of a gate on the die. This allows the correlation to be expressed as a continuous function of the location of a gate. In addition, the authors show that KLE provides much greater accuracy as compared to PCA of a grid model [29], or equivalently, it provides similar accuracy with a reduction in the number of RVs.

2) Propagation of Delay Dependence on Parameter Variations: All the described correlation models share the common characteristic that the set of correlated device parameters is represented by a (typically linear) function of independent RVs. In block-based SSTA, this representation of the correlation is then carried over to the delay and signal arrival times, which are represented in a so-called canonical form. In this section, we will assume that the canonical form is a linear function of normal RVs, which allows us to express the canonical form as

$$d_a = \mu_a + \sum_{i}^{n} a_i z_i + a_{n+1} R \tag{5}$$

where μ_a is the mean delay, z_i represents the n independent RVs used to express the spatially correlated device-parameter variations, R represents the residual independent variation, and coefficients $a_i's$ represent the sensitivity of delay to each of the RVs. The crucial step is to express the results of both the sum and maximum operations in canonical form. This allows the expression of the arrival time to be maintained in canonical form during propagation through the timing graph. This in turn enables the use of a single sum and maximum operation at all locations in the timing graph.

The first operation requires the computation of the sum C of two delay distributions A and B, $C = A + B$, where A, B, and C are expressed in canonical form. Due to the nature of the sum operation, C can be easily expressed in canonical form, and its coefficients can be computed as

$$\mu_c = \mu_a + \mu_b \tag{6}$$

$$c_i = a_i + b_i \qquad \forall i : 1 \le i \le n \tag{7}$$

$$c_{n+1} = \sqrt{a_{n+1}^2 + b_{n+1}^2}. \tag{8}$$

Note that since the last term represents independent variations, it is not correlated with the canonical expressions for A and B. The overall contribution of these independent variations to C is therefore obtained by computing the root sum square of the individual independent contributions.

The second operation requires the computation of $C = \text{maximum}(A, B)$. Since the maximum is a nonlinear function, the maximum of two canonical forms cannot be expressed exactly in canonical form. Hence, the authors in [27] and [29] propose the following algorithm for computing a statistical approximation C_{approx} of the maximum of two arrival times A and B.

1) Compute variances and covariance of A and B

$$\sigma_a^2 = \sum_{i=1}^{n+1} a_i^2, \quad \sigma_b^2 = \sum_{i=1}^{n+1} b_i^2, \quad r = \sum_{i=1}^{n} a_i b_i. \tag{9}$$

2) Compute tightness probability $T_A = P(A > B)$ (the probability that arrival time A is larger than B) as presented in [36]

$$T_A = \Phi\left(\frac{\mu_a - \mu_b}{\theta}\right) \tag{10}$$

where

$$\Phi(x) = \int_{-\infty}^{x} \phi(x)dx \tag{11}$$

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} \tag{12}$$

$$\theta = \sqrt{\sigma_A^2 + \sigma_B^2 - 2r}. \tag{13}$$

3) Compute mean and variance of $C = \text{maximum}(A, B)$ using the results from Clark's work [36]

$$\mu_c = \mu_a T_A + \mu_b(1 - T_A) + \theta\phi\left(\frac{\mu_a - \mu_b}{\theta}\right) \tag{14}$$

$$\begin{aligned}\sigma_c^2 = &\left(\mu_a + \sigma_a^2\right) T_A + \left(\mu_b + \sigma_b^2\right)(1 - T_A)\\ &+ (\mu_a + \mu_b)\theta\phi\left(\frac{\mu_a - \mu_b}{\theta}\right) - \mu_c^2.\end{aligned} \tag{15}$$

4) Compute sensitivity coefficient c_i of C_{approx} using the tightness probability

$$c_i = a_i T_A + b_i(1 - T_A) \qquad \forall i : 1 \leq i \leq n. \tag{16}$$

5) Compute sensitivity coefficient c_{n+1} of canonical form C_{approx} to make the variance of C_{approx} equal to the variance of $C = \text{maximum}(A, B)$. It was shown in [74] that a valid c_{n+1} always exists as the residue $(\sigma_c^2 - \sum_i^n c_i^2)$ is always greater than or equal to zero.

The approach effectively computes the first two moments of C in aforementioned steps 1)–3) and then approximates the maximum using a canonical form in steps 4)–5). The coefficients associated with z_i's are obtained by computing the sum of the two canonical models weighted by their respective tightness probabilities, whereas the coefficient of the independent term is determined to match the variance of C_{approx} and C. In addition, this approximation was shown to match the correlation of C and the z_i's [29].

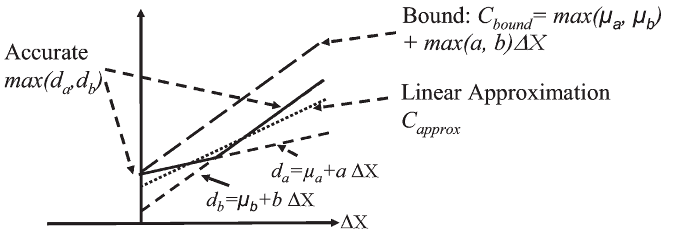
Bound: C_{bound}= max(μ_a, μ_b) + max(a, b)ΔX
Accurate max(d_a,d_b)
Linear Approximation C_{approx}
d_a=μ_a+a ΔX
d_b=μ_b+b ΔX
ΔX

Fig. 15. Exact, approximation and bound of the maximum function of d_a and d_b.

In [28], the authors propose a similar canonical form. However, the independent component of variation is propagated as a discretized delay distribution. The assumption of normal RVs in the canonical form is relaxed, whereas the PDF of the independent variables is assumed to be bounded. The sum operation is performed as mentioned previously. However, the independent term of the sum is obtained by numerically convolving the two independent distributions. The maximum of two PDFs is approximated by computing a bound on the exact maximum using the relation that

$$\max\left(\sum_{i=1}^{n} a_i, \sum_{i=1}^{n} x_i\right) \leq \sum_{i=1}^{n} \max(a_i, x_i). \tag{17}$$

Using the prior inequality and the fact that a numerical maximum of the discrete distribution, as computed in Section IV-A, produces an upper bound, the authors compute the bound of the maximum in canonical form.

Intuitively, we can understand the previous linear approximations of the maximum as follows. Consider the very simple canonical form for two delays $d_a = \mu_a + a\Delta X$ and $d_b = \mu_b + b\Delta X$, where μ_a and μ_b are the mean delays of d_a and d_b, respectively, and a and b are their sensitivities to the common RV ΔX. In Fig. 15, an example of d_a and d_b is shown as a function of ΔX. The maximum of d_a and d_b is the upper envelope of these two intersecting lines, which is a nonlinear function and cannot be expressed exactly by the canonical form. Hence, to represent this maximum, a linear function of ΔX must be constructed that approximates this nonlinear function. The approach presented by Visweswariah *et al.* [27] and Chang and Sapatnekar [29] does so by weighting the sensitivity of d_a and d_b to ΔX by their statistical importance and constructing an approximation labeled c_{approx} in Fig. 15. Note that c_{approx} will at times underestimate and at times overestimate the actual result. On the other hand, the method proposed in [28] constructs a bound $d_{c_{\text{bound}}} = \mu_{c_{\text{bound}}} + c_{\text{bound}}\Delta X$, where $\mu_{c_{\text{bound}}} = \max(\mu_a, \mu_b)$, and $c_{\text{bound}} = \max(a, b)$. As can be seen, the error of c_{approx} will be smaller than that of c_{bound}, whereas c_{bound} will be guaranteed conservative.

Note that the aforementioned methods do not consider the correlation of the statistically independent variation due to reconvergence. Extended canonical models have therefore been proposed by Zhang *et al.* [75], [76], which maintain a separate term for the independent variation associated with each gate. This leads to a significant increase in the size of the canonical form and the ensuing computational complexity. Hence, the authors have also proposed pruning techniques to reduce the size of these canonical expressions.

C. Nonlinear and Nonnormal Approaches

We have discussed various phenomena that result in nonnormal delay and arrival-time distributions. Recently, the problem of nonnormal STA has attracted a lot of attention. Unfortunately, these statistical timing approaches to address nonnormal physical device-parameter distributions or nonlinear delay dependences incur significant computational overheads.

In [31], [32], and [77], the authors extend the linear canonical delay model to include quadratic dependences on device parameters. Following a PCA analysis of the set of correlated device parameters, the canonical form can be expressed as

$$d_a = \mu_a + \sum_{i}^{n} a_i z_i + \sum_{i=1}^{n} \sum_{j=1}^{n} b_{ij} z_i z_j + a_{n+1} R. \tag{18}$$

To enable arrival-time computations using such a quadratic delay model, we need to define the sum and maximum operations using quadratic canonical expressions. The sum of two expressions in quadratic form can be readily seen to result in a quadratic canonical form. In [31], the authors use moment matching to calculate the maximum, again expressed in the quadratic canonical form, where the moments are calculated using numerical integration. The same problem is handled using a conditional linear maximum operation in [32]. The conditional maximum operation implies that the maximum operation is carried out only under the condition that the resulting distribution has a small skew. An estimate of the skewness of the maximum is obtained by assuming the two input distributions to be normal RVs with the same mean, variance, and correlation (as the original RVs) and then using Clark's expression [36] to estimate the skew of the maximum. If the skewness is above the threshold, the maximum is postponed, and both distributions are propagated through the circuit as a set. If the skewness is below a threshold, the maximum is computed at that node. To limit the number of distributions being propagated through each node of the DAG, two distributions are immediately replaced by their maximum as soon as their skewness is found to be lower than the threshold.

Nonnormal physical or electrical device parameters with linear dependences are considered in [34], which uses a canonical expression of the form

$$d_a = \mu_a + \sum_{i}^{n} a_i z_i + \sum_{j=1}^{m} a_{n+j} z_{n+j} + a_{n+1} R \tag{19}$$

where z_1 to z_n represent sources of normal variations, and z_{n+1} to z_{n+m} are RVs with nonnormal variations. The authors use independent component analysis as an analog of PCA to map the correlated nonnormal RVs to a set of uncorrelated RVs. The sum operation using such a canonical expression can be directly expressed in canonical form. The maximum operation is performed using a moment-matching technique based on asymptotic probability extraction [78].

In [30], the authors generalize the first-order canonical form by allowing nonlinear dependences or nonnormal device variations to be included as

$$d_a = \mu_a + \sum_{i}^{n} a_i z_i + f(z_{n+1}, \ldots, z_{n+m}) + a_{n+1} R \tag{20}$$

where f represents the nonlinear function and is described as a table for computational purposes, and RVs z_{n+1} to z_{n+m} represent sources of normal variations with nonlinear dependences or nonnormal variations. The tables describing the nonlinear part of the canonical expressions are computed numerically to perform the sum operation. The maximum operation is performed using tightness probabilities. The tightness probability and the first two moments of the maximum are computed by estimating their value conditioned on the value of the nonlinear parameters and then combining them using Bayes' theorem. The conditional values can be estimated in the same manner as in the linear case described above. The authors show that this approach is efficient for the cases where only a few of the parameters demonstrate complex nonlinear and/or nonnormal behavior.

The SSTA approach presented in [33] uses a Taylor-series expansion-based polynomial representation of gate delays and arrival times, which is able to effectively capture the nonlinear dependences. The sum of such two expressions results in a polynomial representation, thus retaining the canonical form. To compute the maximum the authors use regression analysis while limiting the approximating polynomial representation of the maximum to a reasonable order. In addition, the authors propose to propagate both polynomial and linear delay expressions to reduce the overall complexity of the approach. The linear delay expressions are then used to estimate the tightness probability and the first two moments of the distributions of the maximum. The coefficients of the polynomial approximation of the maximum are then obtained by combining the two original polynomial expressions weighted by their tightness probability and then scaling the coefficients to match the computed first two moments.

In [35], the authors propose the use of skewed normal distributions to model arrival-time distributions. Skewed normal distributions generalize the normal distribution with an additional skewness. This allows for better accuracy when the maximum of two normal RVs results in an output distribution with significant skew. In addition, the authors extend the Clark-based maximum operation to compute a maximum of two skewed normal distributions.

V. Bringing SSTA to Maturity

DSTA has evolved substantially over the last two decades and handles several technology-scaling-related issues such as resistive and inductive shielding, crosstalk noise, power/ground noise, clock jitter and skew, transparent latches, and more. However, most SSTA researchers have, to date, mainly focused on solving the basic SSTA algorithm—the sum and maximum operation required for the propagation of arrival times from the source node to the sink node of the timing graph for combinational circuits. For the adoption of an STA approach, the capabilities of this basic SSTA algorithm must be extended to match the current state of DSTA. For instance, the basic SSTA method presented here assumes that edge weights of all edges in the timing graph are given. Therefore, new methods are required for modeling gate delays, coupling noise, and interconnect delays in the presence of process variations.

Recently, a few approaches have been proposed for addressing these issues in SSTA. A response-surface-method-based statistical gate-delay-modeling technique that handles both die-to-die and within-die variations was presented in [79] and [80]. The authors in [81] propose a statistical gate-delay-modeling technique that considers multiple input switching. In [82], a probabilistic collocation-based method is presented to efficiently construct statistical gate-delay models. An approach to compute statistical gate delays in canonical form is presented using approximate variational $RC-\pi$ load and a variational effective-capacitance calculation procedure in [83] and [84]. A statistical framework for modeling the effect of crosstalk-induced coupling noise on timing was presented in [85] and [86].

Likewise, several approaches have been proposed for performing variational interconnect analysis [87]–[98]. The work in [87] presents a study of the effect of interconnect-parameter variations on projection-based model-order-reduction techniques using matrix-perturbation theory. A balanced-truncation (BTR) method for the analysis of interconnects is proposed in [88] for handling interconnect variations. Using linear fractional transforms, the authors in [89] present a BTR-like method for performing model-order reduction. A method for finding the functional representation of interconnect response in terms of polynomial chaos of process variations is proposed in [90]. Finally, the authors in [91] and [92] have proposed statistical extensions of the empirical D2M delay metric for approximating the distribution of interconnect delay.

STA of sequential circuits is another area that still requires significant investigation. Several issues related to sequential timing, such as accurate modeling of variations and dependences in the clock tree, clock-skew analysis, common-path pessimism-removal problem, and clock-schedule verification for multiple clock domains and latch-based designs, still need to be resolved. Recently, several research efforts have focused on statistical analysis of latch-based designs [21], [23]–[25]. Deterministic timing analysis for latch designs is traditionally performed using the Bellman–Ford algorithm. However, inaccuracies in the maximum function preclude a direct extension of such an approach to statistical timing. In [21], the authors map the timing-analysis problem to a cycle-detection problem on a transformed graph. The graph in the case of early delay analysis (under a conservative formulation [99]) is amenable to such an approach. However, cycle detection requires multiple iterations in the case of late-delay analysis.

Finally, for an SSTA to enable effective circuit-optimization methods, efficient methods for slack computation are needed, which utilize incremental analysis. Some initial approaches for slack computation in SSTA are given in [100]–[102]. Since a more detailed discussion on statistical-optimization methods is not possible in the scope of this paper, the reader is referred to the following publications on this topic [103]–[140].

Apart from these issues pertaining to the timing analysis and optimization methodology, there are other key practical challenges that need to be resolved. Process variability also impacts postfabrication testing, and hence, testing methods must also be extended to find statistical critical paths [141]–[146]. Furthermore, the foundries must devise new test and measurement structures for extracting spatial-correlation models. Similar to design-rule checks, the designers and foundries need to agree on a business model for communicating the statistical characteristics of the fabrication process (i.e., the device-parameter distributions and their correlation) without disclosing proprietary information. Some initial progress is underway in this area. One of the leading semiconductor foundries has recently announced support for statistical-design methodology in its reference flows for 65-nm technology [147].

VI. Conclusion

SSTA has gained extensive interest in recent years. Numerous research findings have been published in literature and a number of commercial efforts are underway at this time. However, the obstacles to widespread adoption of SSTA in industry remain formidable. Much attention has been paid to the modeling and analysis of spatial correlations, nonnormal physical device-parameter distributions, and nonlinear delay dependences. However, the current state-of-the-art SSTA methods still do not address many of the issues that are taken for granted in DSTA, such as interconnect analysis, coupling noise, clocking issues, and complex delay modeling. A number of these issues can probably be addressed fairly rapidly, whereas others may prove more evasive. In addition, a major concern is the lack of silicon verification of the proposed methods and the limited availability of statistical foundry data in a standard format. Finally, SSTA must move beyond pure analysis to optimization to be truly useful for the designer. If silicon shows that statistically optimized designs have substantially higher yield than deterministically optimized designs, an adoption of SSTA is virtually guaranteed.

References

[1] S. Nassif, “Modeling and forecasting of manufacturing variations (embedded tutorial),” in *Proc. ASP-DAC*, 2001, pp. 145–149.

[2] L. Scheffer, “The count of Monte Carlo,” in *Proc. TAU Int. Workshop Timing*, 2004.

[3] T. Kirkpatrick and N. Clark, “PERT as an aid to logic design,” *IBM J. Res. Develop.*, vol. 10, no. 2, pp. 135–141, Mar. 1966.

[4] H. Jyu, S. Malik, S. Devdas, and K. Keutzer, “Statistical timing analysis of combinational logic circuits,” *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 1, no. 2, pp. 126–137, Jun. 1993.

[5] R. Brashear, N. Menezes, C. Oh, L. Pillage, and M. Mercer, “Predicting circuit performance using circuit-level statistical timing analysis,” in *Proc. DATE*, Mar. 1994, pp. 332–337.

[6] S. Das, S. Pant, D. Roberts, and S. Seokwoo, “A self-tuning DVS processor using delay-error detection and correction,” in *Proc. IEEE Symp. VLSI Circuits*, 2005, pp. 258–261.

[7] J. Tschanz, J. Kao, S. Narendra, R. Nair, D. Antoniadis, A. Chandrakasan, and V. De, “Adaptive body bias for reducing impacts of die-to-die and within-die parameter variations on microprocessor frequency and leakage,” *IEEE J. Solid-State Circuits*, vol. 37, no. 11, pp. 1396–1402, Nov. 2002.

[8] G. Nanz and L. Camilletti, “Modeling of chemical-mechanical polishing: A review,” *IEEE Trans. Semicond. Manuf.*, vol. 8, no. 4, pp. 382–389, Nov. 1995.

[9] C. Mack, “Understanding focus effects in submicrometer optical lithography: A review,” *Opt. Eng.*, vol. 32, no. 10, pp. 2350–2362, Oct. 1993.

[10] L. Scheffer, “Physical CAD changes to incorporate design for lithography and manufacturability,” in *Proc. ASP-DAC*, 2004, pp. 768–773.

[11] F. Huebbers, A. Dasdan, and Y. Ismail, “Computation of accurate interconnect process parameter values for performance corners under process variations,” in *Proc. DAC*, 2006, pp. 797–800.

[12] J. Yang, L. Capodieci, and D. Sylvester, “Advanced timing analysis based on post-OPC extraction of critical dimensions,” in *Proc. DAC*, 2005, pp. 359–364.

[13] P. Gupta and F. Heng, “Toward a systematic-variation aware timing methodology,” in *Proc. DAC*, 2004, pp. 321–326.

[14] M. Pelgrom, A. Duinmaijer, and A. Welbers, "Matching properties of MOS transistors," *IEEE J. Solid-State Circuits*, vol. 24, no. 5, pp. 1433–1439, Oct. 1989.

[15] A. Agarwal, D. Blaauw, and V. Zolotov, "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.

[16] A. Agarwal, D. Blaauw, V. Zolotov, and S. Vrudhula, "Statistical timing analysis using bounds and selective enumeration," in *Proc. TAU Int. Workshop Timing*, 2002, pp. 16–21.

[17] J. Kelley, "Critical-path planning and scheduling: Mathematical basis," *J. Oper. Res.*, vol. 9, no. 3, pp. 296–320, May/Jun. 1961.

[18] D. Malcolm, J. Roseboom, C. Clark, and W. Fazar, "Application of a technique for research and development program evaluation," *J. Oper. Res.*, vol. 7, no. 5, pp. 646–669, Sep./Oct. 1959.

[19] J. Hagstrom, "Computational complexity of PERT problems," *Networks*, vol. 18, no. 2, pp. 139–147, 1988.

[20] S. Sapatnekar, *Timing*. New York: Springer-Verlag, 2004.

[21] R. Chen and H. Zhou, "Clock schedule verification under process variations," in *Proc. ICCAD*, 2004, pp. 619–625.

[22] A. Agarwal, D. Blaauw, and V. Zolotov, "Statistical clock skew analysis considering intra-die process variations," in *Proc. ICCAD*, 2003, pp. 914–921.

[23] L. Zhang, Y. Hu, and C. Chen, "Statistical timing analysis in sequential circuit for on-chip global interconnect pipelining," in *Proc. DAC*, 2004, pp. 904–907.

[24] R. Rutenbar, L. Wang, K. Cheng, and S. Kundu, "Static statistical timing analysis for latch-based pipeline designs," in *Proc. ICCAD*, 2004, pp. 468–472.

[25] L. Zhang, J. Tsai, W. Chen, Y. Hu, and C. Chen, "Convergence-provable statistical timing analysis with level-sensitive latches and feedback loops," in *Proc. ASP-DAC*, 2006, pp. 941–946.

[26] C. Visweswariah, "Death, taxes and failing chips," in *Proc. DAC*, 2003, pp. 343–347.

[27] C. Visweswariah, K. Ravindran, K. Kalafala, S. Walker, and S. Narayan, "First-order incremental block-based statistical timing analysis," in *Proc. DAC*, 2004, pp. 331–336.

[28] A. Agarwal, D. Blaauw, and V. Zolotov, "Statistical timing analysis for intra-die process variations with spatial correlations," in *Proc. ICCAD*, 2003, pp. 900–907.

[29] H. Chang and S. Sapatnekar, "Statistical timing analysis considering spatial correlations using a single PERT-like traversal," in *Proc. ICCAD*, 2003, pp. 621–625.

[30] H. Chang, V. Zolotov, S. Narayan, and C. Visweswariah, "Parameterized block-based statistical timing analysis with non-Gaussian parameters, nonlinear delay functions," in *Proc. DAC*, 2005, pp. 71–76.

[31] Y. Zhan, A. Strojwas, X. Li, T. Pileggi, D. Newmark, and M. Sharma, "Correlation-aware statistical timing analysis with non-Gaussian delay distributions," in *Proc. DAC*, 2005, pp. 77–82.

[32] L. Zhang, W. Chen, Y. Hu, J. Gubner, and C. Chen, "Correlation-preserved non-Gaussian statistical timing analysis with quadratic timing model," in *Proc. DAC*, 2005, pp. 83–88.

[33] V. Khandelwal and A. Srivastava, "A general framework for accurate statistical timing analysis considering correlations," in *Proc. DAC*, 2005, pp. 89–94.

[34] J. Singh and S. Sapatnekar, "Statistical timing analysis with correlated non-Gaussian parameters using independent component analysis," in *Proc. DAC*, 2006, pp. 155–160.

[35] K. Chopra, B. Zhai, D. Blaauw, and D. Sylvester, "A new statistical max operation for propagating skewness in statistical timing analysis," in *Proc. ICCAD*, 2006, pp. 237–243.

[36] C. Clark, "The greatest of a finite set of random variables," *J. Oper. Res.*, vol. 9, no. 2, pp. 145–162, Mar./Apr. 1961.

[37] J. Jess, K. Kalafala, S. Naidu, R. Otten, and C. Visweswariah, "Statistical timing for parametric yield prediction of digital integrated circuits," in *Proc. DAC*, 2003, pp. 932–937.

[38] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller, "Equation of state calculations by fast computing machines," *J. Chem. Phys.*, vol. 21, no. 6, pp. 1087–1092, Jun. 1953.

[39] S. Tasiran and A. Demir, "Smart Monte Carlo for yield estimation," in *Proc. TAU Int. Workshop Timing*, 2006.

[40] R. Kanj, R. Joshi, and S. Nassif, "Mixture importance sampling and its application to the analysis of SRAM designs in the presence of rare failure events," in *Proc. DAC*, 2006, pp. 69–72.

[41] V. Veetil, D. Blaauw, and D. Sylvester, "Critically aware latin hypercube sampling for efficient statistical timing analysis," in *Proc. TAU Int. Workshop Timing*, 2007, pp. 24–30.

[42] A. Gattiker, S. Nassif, R. Dinakar, and C. Long, "Timing yield estimation from static timing analysis," in *Proc. ISQED*, 2001, pp. 437–442.

[43] A. Agarwal, D. Blaauw, V. Zolotov, S. Sundareswaran, M. Zhou, K. Gala, and R. Panda, "Statistical delay computation considering spatial correlations," in *Proc. ASP-DAC*, 2003, pp. 271–276.

[44] C. Amin, N. Menezes, K. Killpack, F. Dartu, U. Choudhury, N. Hakim, and Y. Ismail, "Statistical static timing analysis: How simple can we get?" in *Proc. DAC*, 2005, pp. 652–657.

[45] R. Lin and M. Wu, "A new statistical approach to timing analysis of VLSI circuits," in *Proc. Int. Conf. VLSI Des.*, Jan. 1998, pp. 507–513.

[46] B. Choi and D. Walker, "Timing analysis of combinational circuits including capacitive coupling and statistical process variation," in *Proc. Symp. VLSI Test*, 2000, pp. 49–54.

[47] M. Orshansky and K. Keutzer, "A general probabilistic framework for worst case timing analysis," in *Proc. DAC*, 2002, pp. 556–561.

[48] H. Mangassarian and M. Anis, "On statistical timing analysis with inter- and intra-die variations," in *Proc. DATE*, 2005, pp. 132–137.

[49] F. Najm and N. Menezes, "Statistical timing analysis based on a timing yield model," in *Proc. DAC*, 2004, pp. 460–465.

[50] K. Heloue and F. Najm, "Statistical timing analysis with two-sided constraints," in *Proc. ICCAD*, 2005, pp. 829–836.

[51] M. Berkelaar, "Statistical delay calculation, a linear time method," in *Proc. TAU Int. Workshop Timing*, 1997, pp. 15–24.

[52] S. Tsukiyama, M. Tanaka, and M. Fukui, "A statistical static timing analysis considering correlations between delays," in *Proc. ASP-DAC*, 2001, pp. 353–358.

[53] J. Le, X. Li, and L. Pileggi, "STAC: Statistical timing analysis with correlation," in *Proc. DAC*, 2004, pp. 343–348.

[54] K. Kang, B. Paul, and K. Roy, "Statistical timing analysis using levelized covariance propagation," in *Proc. DATE*, 2005, pp. 764–769.

[55] J. Liou, K. Cheng, S. Kundu, and A. Krstic, "Fast statistical timing analysis by probabilistic event propagation," in *Proc. DAC*, 2001, pp. 661–666.

[56] J. Liou, A. Krstic, L. Wang, and K. Cheng, "False-path-aware statistical timing analysis and efficient path selection for delay testing and timing validation," in *Proc. DAC*, 2002, pp. 566–569.

[57] S. Naidu, "Timing yield calculation using an impulse-train approach," in *Proc. ASP-DAC*, 2002, pp. 219–224.

[58] A. Agarwal, D. Blaauw, V. Zolotov, and S. Vrudhula, "Statistical timing analysis using bounds," in *Proc. DATE*, 2003, pp. 62–67.

[59] A. Agarwal, D. Blaauw, V. Zolotov, and S. Vrudhula, "Computation and refinement of statistical bounds on circuit delay," in *Proc. DAC*, 2003, pp. 348–353.

[60] S. Bhardwaj, S. Vrudhula, and D. Blaauw, "τAU: Timing analysis under uncertainty," in *Proc. ICCAD*, 2003, pp. 615–620.

[61] A. Devgan and C. Kashyap, "Block-based static timing analysis with uncertainty," in *Proc. ICCAD*, 2003, pp. 607–614.

[62] V. Khandelwal, A. Davoodi, and A. Srivastava, "Efficient statistical timing analysis through error budgeting," in *Proc. ICCAD*, 2004, pp. 473–477.

[63] R. Topaloglu and A. Orailoglu, "Forward discrete probability propagation method for device performance characterization under process variations," in *Proc. ASP-DAC*, 2005, pp. 220–223.

[64] L. Scheffer, "Explicit computation of performance as a function of process variation," in *Proc. TAU Int. Workshop Timing*, 2002, pp. 1–8.

[65] I. Jolliffe, *Principal Component Analysis*. New York: Springer-Verlag, 2002.

[66] D. Boning and S. Nassif, "Models of process variations in device and interconnect," in *Design of High-Performance Microprocessor Circuits*, A. Chandrakasan, Ed. Piscataway, NJ: IEEE Press, 2000.

[67] Y. Cao and L. Clark, "Mapping statistical process variations toward circuit performance variability: An analytical modeling approach," in *Proc. DAC*, 2005, pp. 658–663.

[68] J. Xiong, V. Zolotov, and L. He, "Robust extraction of spatial correlation," in *Proc. ISPD*, 2006, pp. 2–9.

[69] L. Zhang, J. Shao, and C. Chen, "Non-Gaussian statistical parameter modeling for SSTA with confidence interval analysis," in *Proc. ISPD*, 2006, pp. 33–38.

[70] K. Chopra, N. Shenoy, and D. Blaauw, "Variogram based robust extraction of process variation," in *Proc. TAU Int. Workshop Timing*, 2007, pp. 112–117.

[71] F. Liu, "How to construct spatial correlation models: A mathematical approach," in *Proc. TAU Int. Workshop Timing*, 2007, pp. 106–111.

[72] B. Cline, K. Chopra, and D. Blaauw, "Analysis and modeling of CD variation for statistical static timing," in *Proc. ICCAD*, 2006, pp. 60–66.

[73] S. Bhardwaj, S. Vrudhula, P. Ghanta, and Y. Cao, “Modeling of intra-die process variations for accurate analysis and optimization of nano-scale circuits,” in *Proc. DAC*, 2006, pp. 791–796.

[74] D. Sinha, N. Shenoy, and H. Zhou, “Statistical gate sizing for timing yield optimization,” in *Proc. ICCAD*, 2005, pp. 1037–1041.

[75] L. Zhang, W. Chen, Y. Hu, and C. Chen, “Statistical timing analysis with extended pseudo-canonical timing model,” in *Proc. DATE*, 2005, pp. 952–957.

[76] L. Zhang, Y. Hu, and C. Chen, “Block based statistical timing analysis with extended canonical timing model,” in *Proc. ASP-DAC*, 2005, pp. 250–253.

[77] L. Zhang, Y. Hu, and C. Chen, “Statistical timing analysis with path reconvergence and spatial correlations,” in *Proc. DATE*, 2006, pp. 528–532.

[78] X. Li, J. Le, P. Gopalakrishnan, and L. Pileggi, “Asymptotic probability extraction for non-normal distributions of circuit performance,” in *Proc. ICCAD*, 2004, pp. 2–9.

[79] K. Okada, K. Yamaoka, and H. Onodera, “A statistical gate-delay model considering intra-gate variability,” in *Proc. ICCAD*, 2003, pp. 908–913.

[80] K. Okada, K. Yamaoka, and H. Onodera, “A statistical gate delay model for intra-chip and inter-chip variabilities,” in *Proc. ASP-DAC*, 2003, pp. 31–36.

[81] A. Agarwal, F. Dartu, and D. Blaauw, “Statistical gate delay model considering multiple input switching,” in *Proc. DAC*, 2004, pp. 658–663.

[82] Y. Kumar, J. Li, C. Talarico, and J. Wang, “A probabilistic collocation method based statistical gate delay model considering process variations and multiple input switching,” in *Proc. DATE*, 2005, pp. 770–775.

[83] S. Abbaspour, H. Fatemi, and M. Pedram, “VGTA: Variation-aware gate timing analysis,” in *Proc. ICCD*, 2006, pp. 351–356.

[84] S. Abbaspour, H. Fatemi, and M. Pedram, “Parameterized block-based non-Gaussian statistical gate timing analysis,” in *Proc. ASP-DAC*, 2006, pp. 947–952.

[85] D. Sinha and H. Zhou, “A unified framework for statistical timing analysis with coupling and multiple input switching,” in *Proc. ICCAD*, 2005, pp. 837–843.

[86] M. Agarwal, K. Agarwal, D. Sylvester, and D. Blaauw, “Statistical modeling of cross-coupling effects in VLSI interconnects,” in *Proc. ASP-DAC*, 2005, pp. 503–506.

[87] Y. Liu, L. Pileggi, and A. Strojwas, “Model order-reduction of RC(L) interconnect including variational analysis,” in *Proc. DAC*, 1999, pp. 201–206.

[88] P. Heydari and M. Pedram, “Model reduction of variable-geometry interconnects using variational spectrally-weighted balanced truncation,” in *Proc. ICCAD*, 2001, pp. 586–591.

[89] J. Wang and O. Hafiz, “A linear fractional transform (LFT) based model for interconnect parametric uncertainty,” in *Proc. ISQED*, 2004, pp. 375–380.

[90] J. Wang, P. Ghanta, and S. Vrudhula, “Stochastic analysis of interconnect performance in the presence of process variations,” in *Proc. ICCAD*, 2004, pp. 880–886.

[91] K. Agarwal, D. Sylvester, D. Blaauw, F. Liu, S. Nassif, and S. Vrudhula, “Variational delay metrics for interconnect timing analysis,” in *Proc. DAC*, 2004, pp. 381–384.

[92] P. Ghanta and S. Vrudhula, “Variational interconnect delay metrics for statistical timing analysis,” in *Proc. ISQED*, 2006, pp. 19–24.

[93] R. Jiang, W. Fu, J. Wang, V. Lin, and C. Chen, “Efficient statistical capacitance variability modeling with orthogonal principle factor analysis,” in *Proc. ICCAD*, 2005, pp. 683–690.

[94] X. Li, P. Li, and L. Pileggi, “Parameterized interconnect order reduction with explicit-and-implicit multi-parameter moment matching for inter/intra-die variations,” in *Proc. ICCAD*, 2005, pp. 806–812.

[95] J. Ma and R. Rutenbar, “Interval-valued reduced order statistical interconnect modeling,” in *Proc. ICCAD*, 2004, pp. 460–467.

[96] N. Kankani, V. Agarwal, and J. Wang, “A probabilistic analysis of pipelined global interconnect under process variations,” in *Proc. ASP-DAC*, 2006, pp. 724–729.

[97] K. Yamada and N. Oda, “Statistical corner conditions of interconnect delay (corner LPE specifications),” in *Proc. ASP-DAC*, 2006, pp. 706–711.

[98] S. Abbaspour, H. Fatemi, and M. Pedram, “Non-Gaussian statistical interconnect timing analysis,” in *Proc. DATE*, 2006, pp. 1–6.

[99] N. Shenoy, R. Brayton, and A. Sangiovanni-Vincentelli, “Graph algorithms for clock schedule optimization,” in *Proc. ICCAD*, 1992, pp. 132–136.

[100] K. Chopra, S. Shah, A. Srivastava, D. Blaauw, and D. Sylvester, “Parametric yield maximization using gate sizing based on efficient statistical power and delay gradient computation,” in *Proc. ICCAD*, 2005, pp. 1023–1028.

[101] J. Xiong, V. Zolotov, N. Venkateswaran, and C. Visweswariah, “Criticality computation in parameterized statistical timing,” in *Proc. DAC*, 2006, pp. 63–68.

[102] X. Li, J. Le, M. Celik, and L. Pileggi, “Defining statistical sensitivity for timing optimization of logic circuits with large-scale process and environmental variations,” in *Proc. ICCAD*, 2005, pp. 844–851.

[103] M. Hashimoto and H. Onodera, “A performance optimization method by gate sizing using statistical static timing analysis,” in *Proc. ISPD*, 2000, pp. 111–116.

[104] E. Jacobs and M. Berkelaar, “Gate sizing using a statistical delay model,” in *Proc. DATE*, 2000, pp. 283–290.

[105] A. Srivastava, D. Sylvester, and D. Blaauw, “Statistical optimization of leakage power considering process variations using dual-Vth and sizing,” in *Proc. DAC*, 2004, pp. 773–778.

[106] S. Raj, S. Vrudhula, and J. Wang, “A methodology to improve timing yield in the presence of process variations,” in *Proc. DAC*, 2004, pp. 448–453.

[107] S. Mukhopadhyay, H. Mahmoodi, and K. Roy, “Statistical design and optimization of SRAM cell for yield enhancement,” in *Proc. ICCAD*, 2004, pp. 10–13.

[108] M. Mori, H. Chen, B. Yao, and C. Cheng, “A multiple level network approach for clock skew minimization with process variations,” in *Proc. ASP-DAC*, 2004, pp. 263–268.

[109] S. Chang, C. Hsieh, and K. Wu, “Re-synthesis for delay variation tolerance,” in *Proc. DAC*, 2004, pp. 814–819.

[110] S. Choi, B. Paul, and K. Roy, “Novel sizing algorithm for yield improvement under process variation in nanometer technology,” in *Proc. DAC*, 2004, pp. 454–459.

[111] G. Garcea, N. Meijs, K. Kolk, and R. Otten, “Statistically aware buffer planning,” in *Proc. DATE*, 2004, pp. 1402–1403.

[112] J. Tsai, D. Baik, C. Chen, and K. Saluja, “A yield improvement methodology using pre- and post-silicon statistical clock scheduling,” in *Proc. ICCAD*, 2004, pp. 611–618.

[113] M. Mani, A. Devgan, and M. Orshansky, “An efficient algorithm for statistical minimization of total power under timing yield constraints,” in *Proc. DAC*, 2005, pp. 309–314.

[114] A. Agarwal, K. Chopra, D. Blaauw, and V. Zolotov, “Circuit optimization using statistical static timing analysis,” in *Proc. DAC*, 2005, pp. 321–324.

[115] S. Bhardwaj and S. Vrudhula, “Leakage minimization of nano-scale circuits in the presence of systematic and random variations,” in *Proc. DAC*, 2005, pp. 541–546.

[116] J. Singh, V. Nookala, Z. Luo, and S. Sapatnekar, “Robust gate sizing by geometric programming,” in *Proc. DAC*, 2005, pp. 315–320.

[117] A. Davoodi and A. Srivastava, “Probabilistic dual-Vth leakage optimization under variability,” in *Proc. ISLPED*, 2005, pp. 143–148.

[118] V. Khandelwal, A. Davoodi, A. Nanavati, and A. Srivastava, “A probabilistic approach to buffer insertion,” in *Proc. ICCAD*, 2003, pp. 560–567.

[119] X. Bai, C. Visweswariah, and P. Strenski, “Uncertainty-aware circuit optimization,” in *Proc. DAC*, 2002, pp. 58–63.

[120] A. Singh, M. Mani, and M. Orshansky, “Statistical technology mapping for parametric yield,” in *Proc. ICCAD*, 2005, pp. 511–518.

[121] Y. Zhan, A. Strojwas, M. Sharma, and D. Newmark, “Statistical critical path analysis considering correlations,” in *Proc. ICCAD*, 2005, pp. 699–704.

[122] P. Liu, S. Tan, H. Li, Z. Qi, J. Kong, B. McGaughy, and L. He, “An efficient method for terminal reduction of interconnect circuits considering delay variations,” in *Proc. ICCAD*, 2005, pp. 821–826.

[123] M. Guthaus, N. Venkateswarant, C. Visweswariah, and V. Zolotov, “Gate sizing using incremental parameterized statistical timing analysis,” in *Proc. ICCAD*, 2005, pp. 1029–1036.

[124] J. Xiong, K. Tam, and L. He, “Buffer insertion considering process variation,” in *Proc. DATE*, 2005, pp. 970–975.

[125] A. Datta, S. Bhunia, S. Mukhopadhyay, N. Banerjee, and K. Roy, “Statistical modeling of pipeline delay and design of pipeline under process variation to enhance yield in sub-100 nm technologies,” in *Proc. DATE*, 2005, pp. 926–931.

[126] A. Agarwal, K. Chopra, and D. Blaauw, “Statistical timing based optimization using gate sizing,” in *Proc. DATE*, 2005, pp. 400–405.

[127] O. Neiroukh and X. Song, “Improving the process-variation tolerance of digital circuits using gate sizing and statistical techniques,” in *Proc. DATE*, 2005, pp. 294–299.

[128] D. Sinha and H. Zhou, “Yield driven gate sizing for coupling-noise reduction under uncertainty,” in *Proc. ASP-DAC*, 2005, pp. 192–197.

[129] W. Lam, J. Jam, C. Koh, V. Balakrishnan, and Y. Chen, "Statistical based link insertion for robust clock network design," in *Proc. ICCAD*, 2005, pp. 588–591.
[130] J. Tsai and L. Zhang, "Statistical timing analysis driven post-silicon-tunable clock-tree synthesis," in *Proc. ICCAD*, 2005, pp. 575–581.
[131] G. Venkataraman, C. Sze, and J. Hu, "Skew scheduling and clock routing for improved tolerance to process variations," in *Proc. ASP-DAC*, 2005, pp. 594–599.
[132] W. Lam and C. Koh, "Process variation robust clock tree routing," in *Proc. ASP-DAC*, 2005, pp. 606–611.
[133] L. Deng and M. Wong, "An exact algorithm for the statistical shortest path problem," in *Proc. ASP-DAC*, 2006, pp. 965–970.
[134] M. Guthaus, D. Sylvester, and R. Brown, "Process-induced skew reduction in nominal zero-skew clock trees," in *Proc. ASP-DAC*, 2006, pp. 84–89.
[135] A. Datta, S. Bhunia, J. H. Choi, S. Mukhopadhyay, and K. Roy, "Speed binning aware design methodology to improve profit under parameter variations," in *Proc. ASP-DAC*, 2006, pp. 712–717.
[136] V. Agarwal and J. Wang, "Yield-area optimizations of digital circuits using non-dominated sorting genetic algorithm (yoga)," in *Proc. ASP-DAC*, 2006, pp. 718–723.
[137] M. Ekpanyapong, T. Waterwai, and S. Lim, "Statistical Bellman–Ford algorithm with an application to retiming," in *Proc. ASP-DAC*, 2006, pp. 959–964.
[138] F. Luo, Y. Jia, and W. Dai, "Yield-preferred via insertion based on novel geotopological technology," in *Proc. ASP-DAC*, 2006, pp. 730–735.
[139] D. Patil, S. Yun, S. Kim, A. Cheung, M. Horowitz, and S. Boyd, "A new method for design of robust digital circuits," in *Proc. ISQED*, 2005, pp. 676–681.
[140] A. Davoodi and A. Srivastava, "Probabilistic evaluation of solutions in variability-driven optimization," in *Proc. ISPD*, 2006, pp. 17–24.
[141] J. Liou, A. Krstic, K. Cheng, D. Mukherjee, and S. Kundu, "Performance sensitivity analysis using statistical method and its applications to delay testing," in *Proc. ASP-DAC*, 2000, pp. 587–592.
[142] A. Krstic, L. Wang, K. Cheng, J. Liou, and T. Mak, "Enhancing diagnosis resolution for delay defects based upon statistical timing and statistical fault models," in *Proc. DAC*, 2003, pp. 668–673.
[143] A. Krstic, L. Wang, K. Cheng, J. Liou, and M. Abadir, "Delay defect diagnosis based upon statistical timing models—The first step," in *Proc. DATE*, 2003, pp. 328–333.
[144] Y. Sato, S. Hamada, T. Maeda, A. Takatori, and S. Kajihara, "Evaluation of the statistical delay quality model," in *Proc. ASP-DAC*, 2005, pp. 305–310.
[145] X. Lu, Z. Li, W. Qiu, D. Walker, and W. Shi, "Longest path selection for delay test under process variation," in *Proc. ASP-DAC*, 2004, pp. 98–103.
[146] L. Wang, T. Mak, K. Cheng, and M. Abadir, "On path-based learning and its applications in delay test and diagnosis," in *Proc. DAC*, 2004, pp. 492–497.
[147] Taiwan Semiconductor Manufacturing Corporation, *Reference flow 7.0 release*, 2006. Press Release.

David Blaauw (M'94–SM'07) received the B.S. degree in physics and computer science from Duke University, Durham, NC, in 1986, and the Ph.D. degree in computer science from the University of Illinois, Urbana, in 1991.

Until August 2001, he was with Motorola, Inc., Austin, TX, where he was the Manager of the High Performance Design Technology Group. Since August 2001, he has been joining the faculty of the University of Michigan, Ann Arbor, as an Associate Professor. His work has focused on very-large-scale-integration and computer-aided designs with particular emphasis on circuit design and optimization for high-performance and low-power applications.

Dr. Blaauw was the Technical Program Chair and the General Chair for the International Symposium on Low Power Electronic and Design and was the Technical Program Cochair and member of the Executive Committee of the Association for Computing Machinery/IEEE Design Automation Conference. He is currently a member of the International Solid-State Circuits Conference Technical Program Committee.

Kaviraj Chopra (S'99) received the B.E. degree in instrumentation and control from Gujarat University, Ahmedabad, India, in 2001, and the M.S. degree in electrical and computer engineering from the University of Arizona, Tucson, in 2004. He is currently working toward the Ph.D. degree at the Electrical Engineering and Computer Science Department, University of Michigan, Ann Arbor.

He was with the National Optical and Astronomical Observatory, Tucson, during the winter and summer of 2003. He was a Summer Intern with IBM Corporation, Austin, TX, in 2005, and with Synopsys, Mountain View, CA, in 2006.

Ashish Srivastava received the B.Tech. degree in electrical engineering from the Indian Institute of Technology, Kanpur, India, in 2001, and the M.S. and Ph.D. degrees in electrical engineering from the University of Michigan, Ann Arbor, in 2003 and 2005, respectively.

He is currently with Magma Design Automation, Austin, TX, where he is a Senior Member of the technical staff. He is the author of several papers and one book in his areas of research interest, which include computer-aided-design techniques for the analysis of process variations and optimization of digital circuits.

Dr. Srivastava was a recipient of the Intel Foundation Ph.D. Fellowship for the 2004–2005 academic year.

Lou Scheffer (M'84–SM'06) received the B.S. and M.S. degrees from the California Institute of Technology, Pasadena, in 1974 and 1975, respectively, and the Ph.D. degree from Stanford University, Stanford, CA, in 1984.

He was with Hewlett Packard, Palo Alto, CA, from 1975 to 1981, as a Chip Designer and Computer-Aided-Design (CAD) Tool Developer. In 1981, he was with Valid Logic Systems, San Jose, CA, where he did hardware design, developed a schematic editor, and built an IC layout, routing, and verification system. In 1991, Valid merged with Cadence Design System, San Jose, and since then, he has been working on place and route and floorplanning systems. His research interests include physical design, particularly deep submicrometer effects. He enjoys teaching and has taught courses on CAD for electronics at the University of California, Berkeley, and Stanford University, as well as many tutorials for conferences and industry. He is also interested in Search for Extraterrestrial Intelligence (SETI), where he is the author of several papers and the Coeditor of the book *SETI 2020* (SETI Press, 2003), and serves on the technical advisory board for the Allen Telescope Array at the SETI Institute.