Technische Universität München
Lehrstuhl für Entwurfsautomatisierung

Waveform Based Statistical Timing Analysis of Integrated Digital Circuits

Manuel Volker Schmidt

Vollständiger Abdruck der von der Fakultät für Elektrotechnik und Informationstechnik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktor-Ingenieurs

genehmigten Dissertation.

Vorsitzender: Univ.-Prof. Dr.-Ing. Wolfgang Utschick

Prüfer der Dissertation:

1. Univ.-Prof. Dr.-Ing. Ulf Schlichtmann
2. Univ.-Prof. Dr. rer. nat., Dr. h. c. mult. Manfred Glesner, Technische Universität Darmstadt

Die Dissertation wurde am 30.09.2008 bei der Technischen Universität München eingereicht und durch die Fakultät für Elektrotechnik und Informationstechnik am 01.12.2008 angenommen.

Waveform Based Statistical Timing Analysis of Integrated Digital Circuits

Manuel Schmidt

September 30, 2008

Preface

Any possible calculation can be performed by an algorithm,
running on a computer that provides sufficient time and storage space.
(Alan M. Turing)

This work is a result of my activities as a research assistant at the Institute for Electronic Design Automation. I would firstly like to thank the head of the institute Professor Ulf Schlichtmann, who gave me the opportunity to do my research work at his institute. His support and advice and the friendly and open atmosphere at the institute made this work possible. Professor Ulf Schlichtmann's commitment was invaluable in initiating relations with industry. These relations were vital in order for me to carry out my research with the aim of applicability.

Furthermore, I would like to thank all the colleagues at the institute for the great time I had there and especially the colleagues of the timing group, Christoph Knoth, Bing Li and Walter Schneider. Our discussions and their contributions and friendly attitude helped me immensely.

As this work is closely related to the industry, I would like to thank the involved persons working for Infineon AG. I am especially thankful to Harald Kinzelbach for helping me out with excellent ideas and knowledge about the industrial environment and Klaus Koch for the collaboration regarding the implementation of the proposed method as an in-house tool.

Finally, I want to express my gratitude towards my family and especially towards my fiancée, Ann, for giving me their love and support.

Contents

Chapter 1

Introduction

The history of integrated digital circuits is a history of success as microelectronics pervade the everyday life of modern people. The most important fact for enabling the success of integrated digital circuits is the ability to produce more and more functions on each chip for drastically reduced cost per function. This is only possible by scaling down the dimensions of the basic building blocks – the transistors. With smaller sizes more transistors can be placed on each chip which led to Gordon Moore formulating his famous law in 1965 [Moo65] stating that the number of components on each chip doubles every year. Moore reduced this to every two years in 1975 [Moo75] and this rate is still valid leading to billions of transistors on present chips.

The concept of the Turing Machine [Tur36] is one of the fundaments of machine computation. The Turing Machine is a theoretical concept and was never built as such. It works by reading a character from a tape and depending on this character and the internal state of the machine, the head moves on over the tape, writes to the tape, and the internal state of the machine changes. The Turing Machine is the basis for the theory of Finite State Machines (FSMs). It is stated that every computation can be accomplished by such a Turing Machine and still most fabricated ICs comprise state machines in order to control the computation of the data. One fundamental implication is that the operations are synchronized. The state machine changes states according to a time signal called *clock*. Therefore, the computations of each step have to be finished before the next clock signal and the internal blocks have to meet certain timing constraints.

Figure 1.1 shows a common circuit structure comprising a data path, a clock path, and two registers. The data path performs the data computation and is located between two storage elements or registers. The clock signal is transported

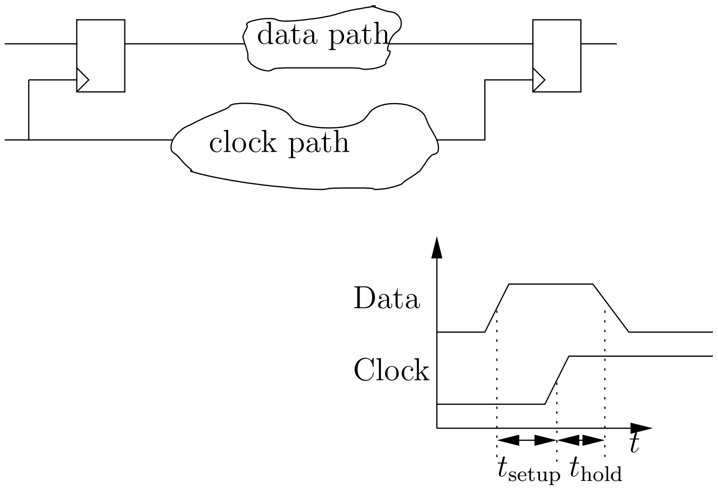

Figure 1.1: Combinatorial path between two flipflops

over the clock path to the receiving register. Due to their physical implementation, the registers exhibit two timing constraints. (i) The data through the combinational data path has to arrive by at least the setup time t_{setup} before the clock edge and (ii) the data signal has to remain stable for at least the hold time t_{hold} after the clock edge. These two constraints ensure the correct latching of the data in the flipflops.

In order to ensure before fabrication that these constraints are met, timing analysis in different stages of the design is crucial. The focus of this work is the timing analysis for semi-custom digital designs. In this design style, logic gates from a given library are used to implement the required logic functions. The transfer characteristics of the single gates have to be modeled in order to obtain sufficiently accurate estimates for the delay of the entire circuit. As the values of the input signals of the circuit are unknown at this stage, this delay estimate has to be independent of the actual signal value assignments, which is called *static* timing analysis (STA).

The shrinking of feature sizes causes accuracy problems in the manufacturing process. The physical structures could never be fabricated with infinite accuracy and with smaller physical dimensions, these imperfections have a larger impact on the timing of the circuit. The International Technology Roadmap For Semiconductors (ITRS) [TIT07] contains some information on the expected development of threshold voltage variation as an example. The values of the estimation are displayed in Figure 1.2. The 3σ value of the variation of the threshold voltage is expected to rise from 17% in 2007 to more than 35% in 2015. According to the ITRS, the need for novel methods of timing analysis considering these variations is evident. Existing methods of STA have to be extended to consider statistical variations of process

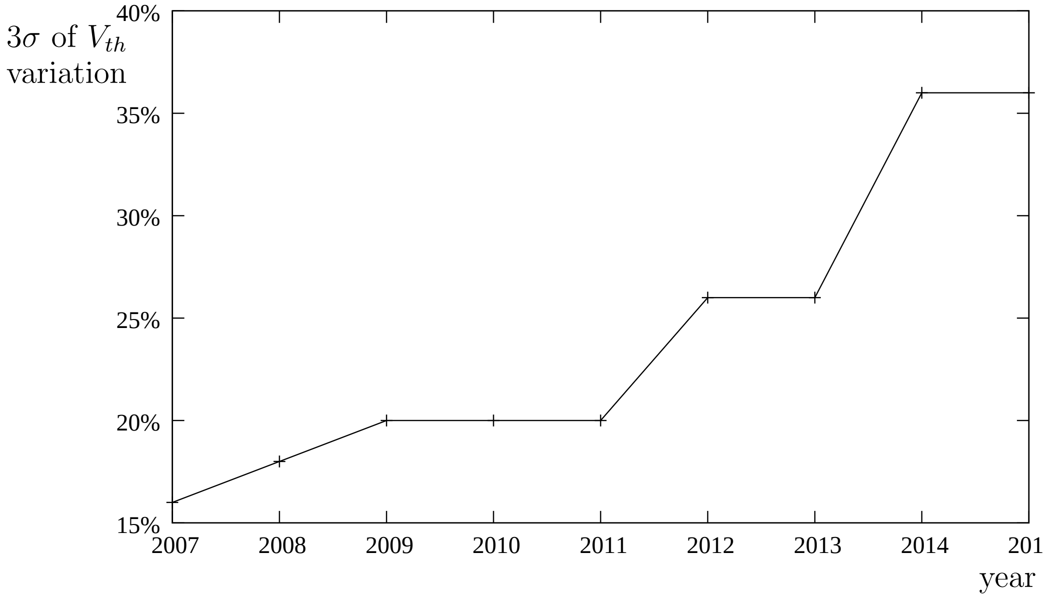

Figure 1.2: Rise of threshold voltage variation.

parameters leading to statistical static timing analysis (SSTA).

1.1 State-of-the-Art

This section gives an overview of the development in the area of timing analysis of digital circuits. It starts from the basic approach to static timing analysis, covers the methods used in industry today and ends with the latest publications of the research community. The first subsection deals with the development of deterministic STA. The second part introduces more process related publications, where the origins, statistical metrology (measurement, characterization, and modeling [SFS+99]) and implications of process variations are discussed. The third part shows the various statistical extensions to deterministic static timing analysis in a mainly chronological order.

1.1.1 Deterministic Static Timing Analysis

Figure 1.3 shows the three main problems which can be identified for timing analysis:

1. Waveform modeling: The waveform has to be described in a way that is compatible with the gate model. Starting with just the arrival time, current industrial tools also model the slope. Methodologies proposed in research literature model the shape of the waveform as well.

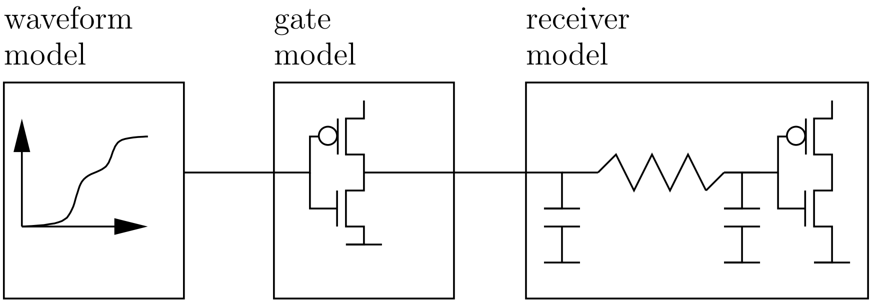

Figure 1.3: Three essential models for timing analysis

2. Gate modeling: The gate model describes the timing characteristics of a single cell considering the input waveform and the output load. The development progressed from a fixed delay to a load dependent delay, further to the consideration of slope and, latest, the shape of the waveform. Since the first linear models proved insufficient, nonlinear analytical models as well as look-up tables were used to capture nonlinear dependencies.

3. Receiver modeling: The receiver or load of a gate consists of the interconnect and all connected gates in the fanout. The entire receiver structure has to be modeled with compatibility to the gate model. The most common modeling is to use a single capacitor, while the value of this capacitor can be computed in various different ways. Only recently, these single capacitances became insufficient and are replaced by more complex structures.

An overview of the literature dealing with these three problems, waveform modeling, gate modeling, and receiver modeling is given in the following. In the beginning of static timing analysis, only the arrival time is propagated through the circuit in a path-based or block-based manner. Fall delays and rise delays are considered separately but the impact of different slopes or loads is neglected. The author of [Hit82] describes how the delay of combinational blocks between memory elements must meet certain timing constraints. The delay must not be too large or too small, otherwise the resulting data can not be successfully captured by the memory elements at the end of the combinational block. Further, the two basic procedures for traversing a circuit are described: Path enumeration and block-based analysis.

Path enumeration works by starting from a particular start point and traversing the circuit backwards until a primary input or other terminal node is reached. This method is very accurate and can detect and eliminate paths which can never be sensitized. However, this method suffers from the high number of possible paths through a circuit and thus, high computational effort.

The second option, block-based traversal, starts at primary inputs or the output of memory elements. All elements to which this starting point is connected to by signals are processed and for each element the earliest and latest arrival times for the output signal is computed. Doing so, the output arrival time of each element is computed only once. The way of proceeding through all elements is adopted from the *Project Evaluation and Review Technique* (PERT). The block-based method is significantly faster than the path-based approach but tends to be pessimistic. Because the logic function of the gates are neglected, specific paths cannot be excluded from the analysis. Thus, paths which would not affect the arrival time at the output – as these paths are not sensitized – would influence the outcome of the analysis possibly leading to a later arrival time.

In order to incorporate the load of the connected cells in the fanout, analytical models were proposed. According to [Sap04] , one of the first approaches models the gate delay as a linear function of the purely capacitive output load: $D = k_1 C_L + k_2$ with gate delay D and output capacitance C_L. Due to the factor k in the linear expression, this model and later models enhancing this basic analytical delay model are called *k-factor models*. One of these enhancements was introduced in [HJ87] where the authors proposed a method to also incorporate the dependence of the input slope to the gate delay. The authors show an analytical solution for the CMOS inverter output response to an input voltage ramp instead of the step response. The delay of a gate is expressed by the step response delay plus a correction term. This correction term is linearly dependent on the input transition time. Hence, the influence of the input transition time can be considered in the computation of the delay of the inverter [WE93].

A major problem for this delay model is the nonlinearity of the relationship between input transition time, load capacitance and delay. To capture this nonlinearity, an empirical approach evolved, which uses look-up tables. The values for gate delay and output transition time are stored dependent on load capacitance and input transition time. This delay model is called *Nonlinear Delay Model (NLDM)* as the nonlinear functions are represented by the look-up tables. The table structure is depicted in Figure 1.4. The entries of the tables as well as the input capacitance of each library cell are obtained during library characterization.

Using these tables, a combinational circuit consisting of cells from the previously characterized library can be analyzed with high efficiency. Starting at a primary input the input transition time is known as the input signal is known. The output

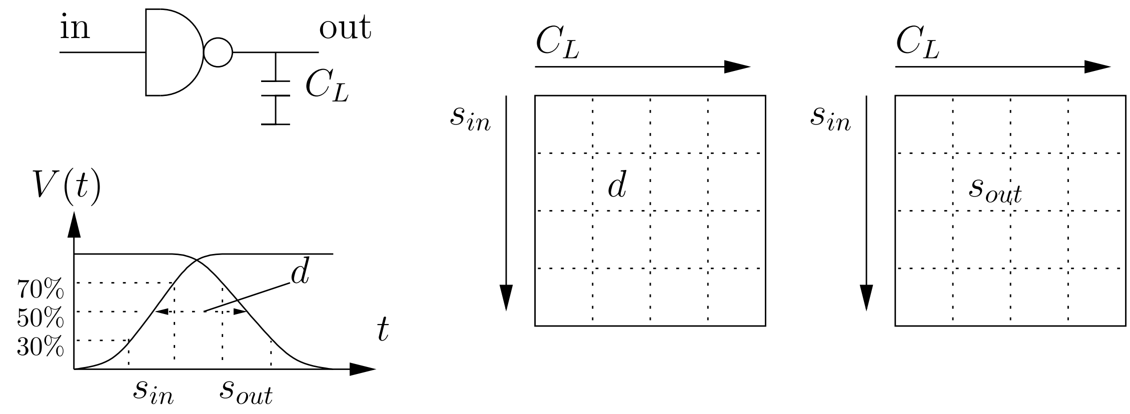

Figure 1.4: Look-up tables for deterministic STA. The tables store delay d and output slope s_{out} depending on input slope s_{in} and output load C_L

load is composed of the input capacitances of the receiving gates and the equivalent capacitance of the interconnect. With this data, the values for the gate delay and the output transition time can be read from the tables. The delay is added to the arrival time at the input and the result is the arrival time at the output. Thus, the previously described block-based traversal of the circuit is extended by the dependency of the input slope to the gate delay and the slope of the signals are propagated through the circuit together with the arrival time.

With scaling down feature sizes, increasing interconnect resistances and decreasing gate capacitances lead to inaccurate results when assuming purely capacitive load at the output [DMQP94]. However, the complex interconnect structures are reduced to a single effective load capacitance in order to use the existing analytical or table-based methods. In [QPP94] the authors address this problem and describe how the resistance of interconnects become more relevant as the length of connected lines does not scale down with feature sizes as the density is growing. Once the wire resistance reaches the values of gate resistance, a significant part of the load capacitance is shielded from the driving gate. Thus, the delay of the driving gate will be smaller as the gate can not “see” the full capacitance. As a solution to this problem the authors propose the incorporation of a resistance model into the computation of the effective capacitance. Mapping the effects of resistive shielding to the effective capacitance allows the further usage of k-factor models.

The effects of resistive shielding on arrival time and slope is captured but the underlying mechanism is still not addressed: Due to the increased resistance of the interconnect, the signal waveforms show long “tails” at the end of a signal transition. Such waveforms differ significantly from the ramp based model as it is used during propagating slope and arrival time. With shrinking feature sizes to the nanometer region, however, this deviation from the ramp based model induces unacceptable inaccuracies to the results of timing analysis [KL01]. It is crucial to refine the modeling of the waveforms and thus to consider a wider range of different waveform shapes. Besides the inaccuracy induced by overly simplistic driver models, a second effect concerning the receivers gained importance. The input capacitance of a gate shows increasingly nonlinear behavior. The main cause for that is the Miller effect. The gate capacitance of the transistors depends on the drain voltage and as the drain voltage changes during the switching of the cell, the input capacitance changes as well.

An enhancement is needed which models the waveform more accurately and considers nonlinear effects in the receiver model but is also compatible to established NLDM in order to be applicable in the industry without changing the entire timing concept. Two slightly different extensions of the NLDM were introduced as *Composite Current Source Model (CCSM)* and *Effective Current Source Model (ECSM)*. ECSM was first introduced by [KL01] proposing to replace each gate by a current source. The current is modeled by a piecewise linear function with one turning point at V_{th} of the output voltage separating the saturation mode from the linear mode of the transistors. The gate model consists of the current source parallel to a linear resistor and parallel to an internal capacitance for small load capacitances. The capacitance values are chosen to match library timing data at any operating point in the look-up tables depending on s_{in} and C_L.

The meaning of ECSM changed slightly as a new ECSM was developed [Kez06]. Load-slope look-up tables are still in use but now for each combination of load and slope a time-voltage waveform is stored as depicted in Figure 1.5. During STA, for a specific effective output load and an input slope the time-voltage waveform is retrieved from the tables. This waveform is then converted to a current waveform and applied to the complex interconnect structure. Solving this system yields the arrival time and slope at the input of the next gate. In order to account for the Miller effect, the receiver gates are modeled by a variable capacitor.

The second extension to NLDM is CCSM which differs only slightly from ECSM. Instead of time-voltage waveform, CCSM stores time-current waveforms. This results in a different library characterization but the general STA procedure is equal to the procedure using ECSM.

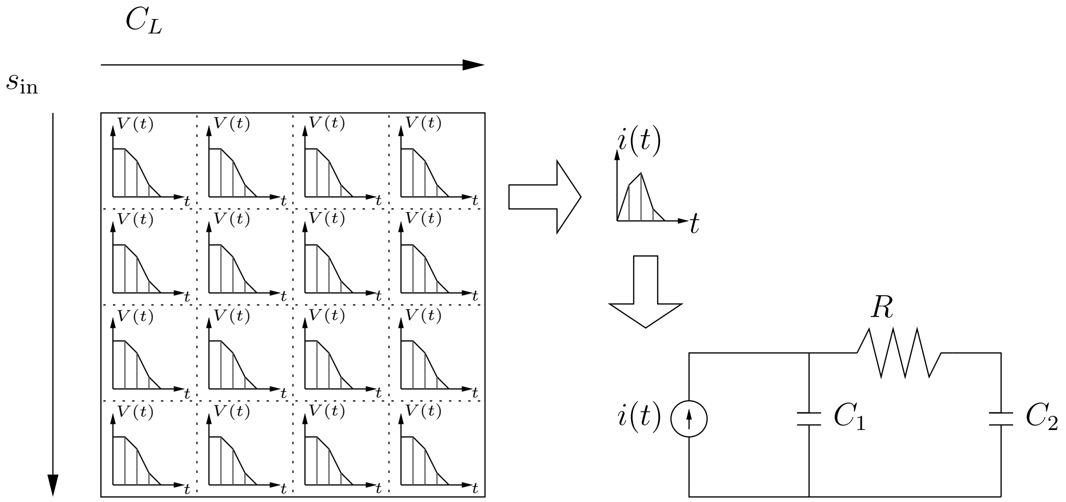

Figure 1.5: Look-up tables for different waveforms depending on input slope and output load and π-model driven by a current source.

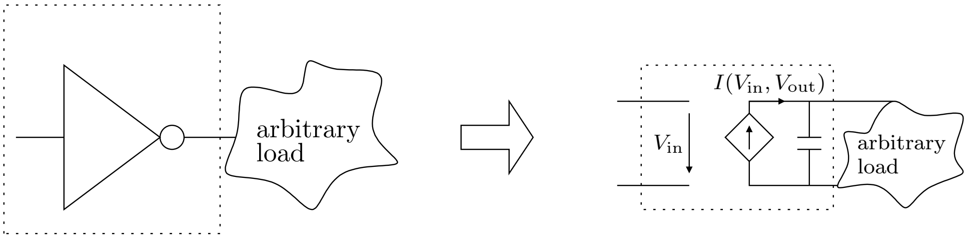

Figure 1.6: Current source model comprising a voltage controlled current source and intrinsic capacitance

In [HYO04] the authors introduce equivalent waveforms. The equivalent waveform is a waveform with a different arrival time than the original waveform but with a standard shape. This equivalent waveform produces the same output waveform as the original input waveform. The use of only standard shapes simplifies the consideration of waveform dependencies. This approach is not used widely by research or industry.

Another approach with different modeling and without compatibility to NLDM is the *Current Source Model (CSM)* (see Figure 1.6). The output current is still stored in look-up tables and instead of current waveforms, the static current depending on V_{in} and V_{out} is stored. These values are obtained by a DC sweep during library characterization. The input signal for one cell is known as well as the initial value of the output voltage. For each time step of the input transition, the current into the receiver model can be retrieved from the look-up table. Solving the appropriate differential equation yields the voltage waveform at the input of the following gate.

The different approaches for CSM differ in the additional elements in the gate model. The authors in [CW03] propose to use a voltage-controlled current source (VCCS) and a constant, intrinsic capacitance. This intrinsic capacitance is used to reflect the effects of parasitic capacitances in the gate. The capacitance value is determined by matching the output signal of the model to the output signal of a transient simulation on transistor level.

In [KTV04] the CSM is enhanced by capturing the Miller effect, using an additional capacitance between the input and output nodes of a gate. Introducing crosstalk into current source models was shown in [KTV04] and also the more detailed modeling of the receiver. A further improvement was described in [FNP06] using an additional capacitance between the input node and ground and in [NP06] the authors describe a CSM method not storing the current in the tables but the derivatives of the current with respect to time. These values are dependent on input voltage and output load. Using these values the output current is progressively computed by numerical integration. The authors state that their model is superior as it reflects the effects of the parasitic capacitances more accurately.

Apart from the current source models, the authors in [ADI03, ADI05] model the waveform using the Weibull function. It is an exponential expression with two parameters resembling typical waveforms. The two parameters can be interpreted as slope and shape parameters. Look-up tables are used to propagate the Weibull parameters through the circuit.
Another timing methodology based on HSPICE simulation of individual cells was proposed in [CM06]. Not the entire waveform but only the delay and the transition time are propagated.

1.1.2 Technology Data

The aim of all the methods described above is to analyze the timing properties of a given circuit for a set of fixed device parameters. However, these parameters deviate from their intended values due to manufacturing imperfections. This section gives an overview of the published literature on the topic of technology data and manufacturing variations.

Variation can be classified into global and local variations. Global variations affect each device on one die equally, i. e. all devices on one die have the same parameter values but different values to devices on other dies. Local variations affect each device differently.

In the design of digital circuits, global variations can be considered by corner case analysis. The set of parameters which causes the worst timing result, e.g. the largest delay of a data path, is determined and the design is tweaked until the timing constraints are met assuming that worst-case corner. As this approach is relatively uncomplicated, the focus of the following is more on local variations.

In contrast to digital circuits, where the influence of absolute values of parameters are dominant, in analog circuit design the values relative to others are more significant than their absolute values. One example is the ratio of the widths of two transistors in a current mirror. The deviation from the intended ratio is called mismatch. These relative values are not affected by global, i.e. perfectly correlated, variations but by local variations. Therefore, local variations became an issue in analog circuit design much earlier than in digital circuit design.

One example for a publication from the analog domain dealing with mismatch is [LHC86]. The authors present measurement data from a 3μm CMOS process with excluded influence of global variation showing more than 11% relative standard deviation $\sigma_{V_{th}}/V_{th}$ of the threshold voltage of a p-channel MOS transistor. This value is for the smallest transistor investigated and it is shown that the relative standard deviation depends linearly on $1/\sqrt{LW}$, where L and W is the transistor length and width respectively.

The influence of local variations on the delay of digital paths was examined in [EBSLM97]. The influence of threshold voltage variation on gate delay was determined for a 0.5μm CMOS process resulting in a relative standard deviation of up to 5%. This value falls below 1% for larger transistors and higher ratio of supply voltage and threshold voltage V_{DD}/V_{th}. The effect on the path delay was investigated by a test structure of a 24 bit carry select adder. The relative standard deviation of the delay of a path of four gates was measured as up to 10% for a 0.5μm process using the lowest threshold voltage and the smallest transistors. This value rose up to 15% for a 0.35μm process. The projection of gate delay variation shows the rise from 5% for a 0.5μm CMOS process up about 15% for a 0.18μm process. Further, it was shown that local variations have most impact on designs with many critical paths with small logic depth, e.g. in highly pipelined circuits.

Also the author in [Nas00, Nas01] shows the trends and sources of process vari-

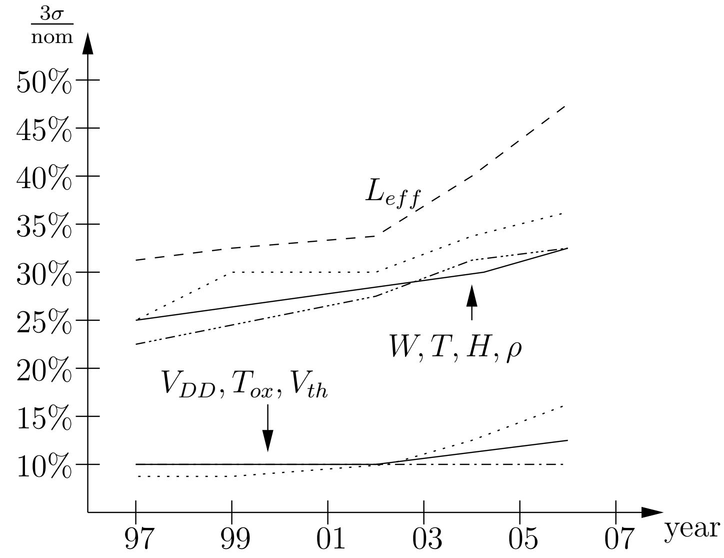

Figure 1.7: Technology parameter variation [Nas00] for device parameters (L_{eff}, T_{ox}, V_{th}) and wire parameters (W, T, H, ρ)

ability and how intra-die variations gain importance in the future. The author shows numbers for the 3σ value relative to the nominal value. The variation of V_{th} is shown to rise from 10% in 1997 for 250nm to 15% in 2006 for 70nm. Besides this moderate increase, the development of the variation of other parameters is considered being more dramatical. The variation of the effective channel length rises from 30% to 45% in the same time range. The contribution of devices (34%) and wires (66%) to the overall delay variation is specified as well as the impact of device and wire parameters on delay. It is shown that the wire parameters have a significant share of the overall delay variation of 66%.

In [BKN$^+$03] the authors present measurement data showing the larger variability in operating frequency of manufactured chips due to process variations. The 3σ value of the V_{th} variation is stated as 30mV causing a 30% variation in the chip operating frequency. It is also shown that lowering V_{th} and the logic depth increases the performance but also increases the variability of the delay of the entire chip. The authors in [OMC$^+$00, OMC$^+$02] show measurements from a 180nm process revealing a large intrachip L_{gate} variation of 10% in average causing a variation of the delay of the critical path of up to 17%. The authors claim further, that the variation of the gate length is rather systematic spatial than random. Thus, the actual value of the gate length depends on the location of the gate on the die and the proposed method to consider this effect is to use different values of the gate length for each gate depending on the location after placement.

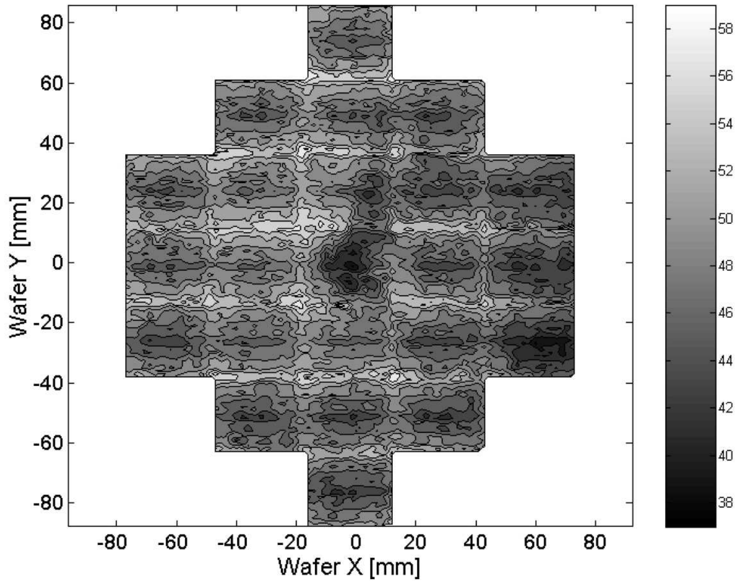

Figure 1.8: A CD contour map as the average over all measured wafers [CS03a]

[CS03a] shows measurements of critical dimension (CD) done by electrical measurements or scanning electron microscopy. The different scales of systematic variation is described. The given CD contour map of measured data is reprinted in Figure 1.8. The contour map is obtained by computing the average of the measurements on all wafers at the positions on the map. The systematic variation of the CD can be seen by the spatial dependence inside each die and also over the entire wafer.

Similar countour maps for a 130nm process are given in [FCC+05] and the problem of modeling spatial correlations is addressed. The authors propose a piecewise linear function of the correlation coefficient depending on the distance. [Kuh07] provides some data of variations of the threshold voltage down to 45nm stating an increase of the standard variation σ from 25mV for 130nm to 45mV for 45nm. The author also proposes a way of simulating the effect of random dopant fluctuation on the threshold voltage. He uses a three-dimensional numerical model with an adaptive local meshing scheme. The results showed a significant deviation from the measurements indicating room for process improvement.

The effect of non-rectangular gate (NRG) for a 65nm process is discussed in [SBS+07]. NRG is caused by the distortion of the light as the optical wavelength of 157nm is much larger than the minimum feature size. The authors point out that this effect mainly influences the off-current of the transistor which can deviate up to 20X. This causes a large deviation of the leakage current. The on-current on the other hand is only deviated by up to 5%. In [YLNC08] the authors mention the increasing impact of NRG on the threshold voltage and describe the integration of this effect together with line edge roughness and random dopant fluctuation into a standard SPICE environment avoiding complex atomistic simulations.

Possible improvements to deal with the increasing influence of process variations are classified in [BBC$^+$08, BBC$^+$07]. The authors note that the basis of improvements is the statistical metrology. This measurement and modeling of process variations generates data which is crucial for the further operations. The metrology data can be handed over in two different directions: The process and equipment control on one hand and to the circuit design in the other hand. The latter aims for a design which is robust to process variations called design for manufacturability (DFM). “A core need is variation impact analysis, particularly tools that can utilize richer representations of process variations, such as statistical timing analysis with spatially correlated process variations” [BBC$^+$08]. The state of the art of this topic will be the focus of the next section.

1.1.3 Statistical Statistic Timing Analysis

Due to unavoidable variations in the fabrication process, the parameters of the circuit elements deviate from their nominal value. Most of these deviations are random variations which can be described by a probability density function (pdf). The traditional way of dealing with these variations is to model the pdfs by intervals formed by the extreme points of the distribution. The points in the process parameters space which lie on the extreme value of every parameter are referred to as corners. The worst-case corner shows the worst combination of extreme points. Therefore, it used to be sufficient to show that the circuit works for the worst-case corner in order to show that it works for all other corners.

For long data paths between two registers, the slowest case for the data path and the fastest case for the clock path is the worst case as a setup time violation can be caused if the data signal does not arrive at the receiving register early enough before the clock signal. For short data paths on the other hand, the worst case is the fastest case for the data path and the slowest case for the clock path as a hold time violation occurs if the data signal of the next clock cycle arrives at the receiving register before the previous data has been latched properly. Regardless wether it is the slowest or the fastest case, the most problematic corner is always referred to as worst-case. The process parameters will not exceed the worst-case parameter set and thus all chips will meet the timing specifications. This concept works well if the process parameters on one die are perfectly correlated, i.e. all devices are in the worst-case corner if one device is in the worst-case corner.

However, with increasing relevance of within-die (i.e. uncorrelated, local) process parameter variations, the delay of the worst-case corner will be overestimated which can be overly pessimistic [SSB05] if the worst-case is the slowest case or even optimistic if the worst-case is the fastest case. If the worst-case is the slowest case, meeting the constraints at the overly pessimistic worst-case corner implies spending a substantial amount of time and area and loosing power efficiency without much benefit. If it is vice versa, the timing constraints could be violated even though the timing check passed. In addition to that, the number of corners in the process parameter space rises exponentially with the number of process parameters. Therefore, it becomes increasingly difficult to determine the worst-case corner.

First attempts to reduce pessimism is the consideration of within-die variations by an on chip variation factor (OCV). The OCV contributes to the fact that the influence of local variations declines for longer paths. Thus, different factors are multiplied on path delays depending on the length of the path. It became clear, however, that process variations need to be considered statistically and that the traditional concepts of timing analysis had to be changed to incorporate the statistics of the parameters. This leads to statistical static timing analysis (SSTA). The random variation of process parameters leads to a variation of gate delays and further to variation of signal parameters such as arrival time and slope. These random variations can be described by pdfs.

First publications on SSTA model gate delays as discrete pdfs [LCKK01, Nai02] or piecewise linear pdfs [DK03]. The statistical information has to be propagated through the circuit. Therefore, two basic operations need to be executed: The arrival time at the input of a cell has to be added to the delay of the cell and as both quantities are random variables, a statistical *add* function has to be available. While the *add* function is straightforward, the second operation can become cumbersome: If a gate has more than one input, the statistical maximum has to be computed by a *max* function. The problem arises if the two input signals are not independent because they share common gates in their path or the process parameters of the gates in the paths are not independent. The problem with dependency is the computational complexity: Each combination of samples of the pdfs on which the two input signals depend has to be considered. The number of combinations rises

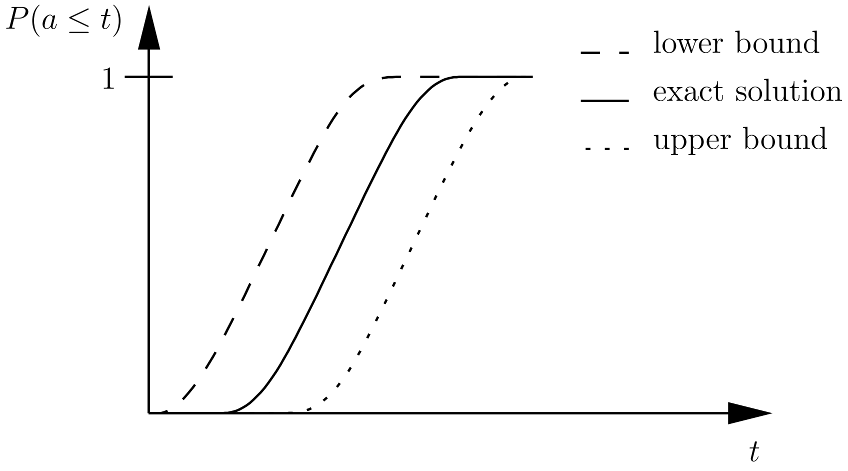

Figure 1.9: Upper and lower bounds of the cdf of the arrival time a

exponentially in the number of dependent pdfs, e.g. common gates of two paths, causing prohibitive computational complexity.

Therefore, instead of computing the exact pdfs, bounds were introduced and it was proven that these bounds are never exceeded. [OK02] describes how to derive the joint probability distribution of the delay of a number of paths. The authors describe a way to obtain upper and lower bounds for the distribution of the maximum of the path delays in order to get an estimate on the circuit delay distribution. [OB04] and [WO06] show how to obtain bounds for N most critical paths using stochastic majorization. In [ABZV03b] the authors propose a block-based method to obtain bounds on the circuit delay. The computation of these bounds is based on neglecting some of the inputs or assuming independence between the arrival times. In [ABZV03a, AZB03] the same method as before is used for obtaining the bounds which are then refined by selective enumeration, i.e. the exact computation of some most significant nodes. The use of Bayesian networks is proposed in [BVB05] in order to obtain bounds for the pdf of the circuit delay. The authors in [JKN$^+$03, JKN$^+$06] show how to find the yield, i.e. the probability that the timing constraints are met by three different ways: 1) Dividing the feasibility region, which is the area in the process parameter space which satisfies timing constraints, into rectangles and integrate the jpdf inside the rectangles, 2) finding the maximal ellipsoid inside the feasibility region and integrate over the volume of that ellipsoid and 3) using tightness probability as in [VRK$^+$04] (see below).

On the one hand, it seemed too complex to compute the exact pdf of a circuit and on the other hand, it was observed, that most process variations can be modeled by Gaussian distributions with sufficient accuracy. As Gaussian distributions are fully described by the mean value μ and the standard deviation σ, it is sufficient to consider only these two characteristics of the random variables. In addition, it was assumed to be sufficient to model the gate delay as a linear function of transistor parameters. This is used in [GNDL01] in a path-based approached, where the n most critical paths are obtained from STA and the sensitivities of the path delay with respect to the process parameter variations are obtained. With this information and assuming Gaussian random variables, the distribution of the path delay can be computed. The linear assumption leads to a simple representation of the arrival times in a circuit by weighted sums of process parameter variations referred to as *canonical sums*. In [ABZ03] this form is used and a method is proposed to compute the addition and bounded maximum operation of two of these canonical sums in order to propagate the canonical sums through the circuit. [CS03b] describes how an estimate of the maximum of two canonical sums can be computed using the work of [Cla61]. In [OYO03] the authors describe in more detail how to obtain the sensitivities for all relevant values of input slope and output load. The representation of the transfer function of an interconnect structure in a canonical form is shown in [AFP06] and [AFP07].

Using the canonical form solves the problem of reconvergent fanouts: The representation as a sum of random variables keeps track of all variables from previous gates. If two signals reconverge, the canonical sums of their arrival times show common variables. With this information the correlation between the two arrival times can be obtained and used for the computation of the maximum. Besides the problem of reconvergent fanouts, the problem of correlated process parameters still remains open. The random variables used in the canonical form have to be independent. If the process parameters are correlated, i. e. not independent, these parameters have to be transformed into a set of independent variables.

Principle Component Analysis (PCA) can be used to decorrelate random variables. For Gaussian random variables uncorrelatedness also implies independence (see Section 2.3.4). The PCA itself is described in [HKO01] while the application to obtain a linear sum of process parameters is shown in [CS03b, CS05b]. A more detailed description of the application to die-to-die and wafer-to-wafer variations and the description of a constraint PCA can be found in [CKK$^+$08].

For non-Gaussian random variables, PCA can not be used as uncorrelatedness does not imply independence. A more complex transformation of a set of random variables into a set of independent random variables is Independent Component

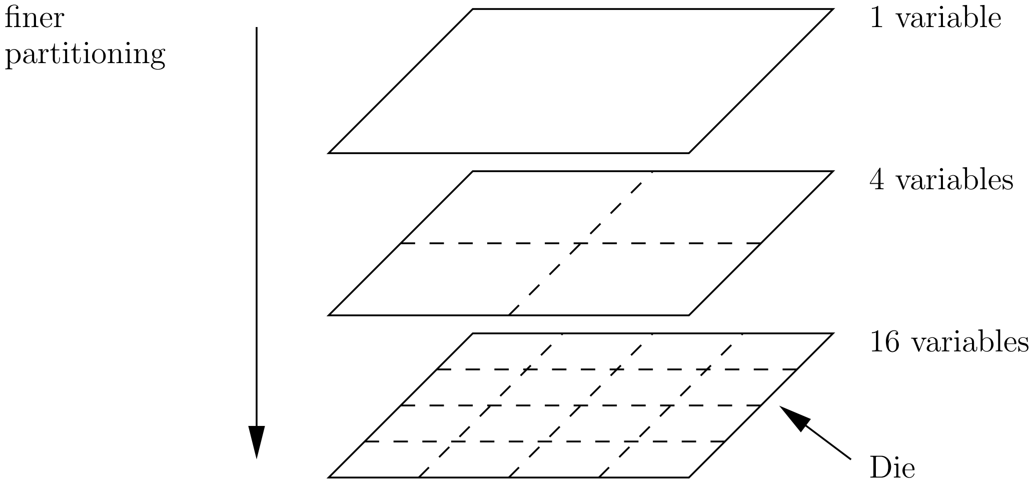

Figure 1.10: Quad-tree partitioning of the die to model spatial correlations [ABZ03].

Analysis (ICA), which is also described in [HKO01] and applied in [SS06] and [SS08]. ICA was developed to separate different sources, e. g. audio signals, from a linear mixture. The basic principle is the reduction of the Gaussianity of the output signals because a mixture of two random variables is always more Gaussian than the original signals. ICA can be applied only on non-Gaussian random variables.

Apart from the transformation of the original random variables into independent variables, other methods exist to obtain a sum of independent random variables more directly. One such method for obtaining a canonical form representing correlated delays is a quad-tree model [ABZ03]. Figure 1.10 shows how several layers are used to partition the die into rectangles with lower layers showing a finer granularity of the partitioning. One independent random variable is assigned to each rectangle. Each parameter of a device on the die is represented by a sum of the random variables which are assigned to the rectangles in which this particular device is located.

In a similar way the die can be divided into a grid [CS03b, CS05b, CS05a, KS05, SSA$^+$05]. Besides quad-tree and grid-based models, correlations can also be modeled by distance-dependent functions. In [FCC$^+$05] a linear function was proposed, but in [LWA06, XZH07a, XZH07b] it is shown that the resulting assumed covariance matrix is not positive semidefinite and thus lacks the property of a covariance matrix. Thus, a methodology is proposed to obtain a nonlinear spatial correlation function and a correlation matrix from measurement data. In [LTCC08] a set of polynomials is used as spatial correlation function using Singular Value Decomposition and Polynomial Fitting. The property of the correlation matrix being positive semidefinite is obtained by a post-processing step. To overcome the problem of modeling the spatial correlation, the authors in [ISNM08] propose a Monte Carlo based approach. In their work, a vector of sampled delays for each gate is used for SSTA.

Using the linear canonical form, the addition of two arrival times or delays is a straightforward addition of the appropriate coefficients. For the maximum operation some of the authors use the analytical expression of Clark [Cla61] to compute the moments of the maximum. A simpler approach to estimate the maximum of two arrival times is to linearize the max function. In [VRK$^+$04, VRK$^+$06] the authors introduce the idea of tightness probability and use it as factors of the canonical form. The tightness probability is defined as the probability that one signal arrives earlier than the other signal: $P(a_1 < a_2)$. Wherein a_1 and a_2 are the arrival times of the two signals. The linear expression for the arrival time o at the gate output is then: $o = P(a_2 < a_1)a_1 + P(a_1 < a_2)a_2$. An extension to this linear approach is presented in [ZCHpC06] where the authors propose to estimate the nonlinearity of the max function at the current gate and keep a tuple of the input arrival times if this estimate of the nonlinearity is above a certain threshold.

The assumption of linear influence of process parameters on gate delay is discussed controversially and some authors claim that linear modeling is too inaccurate. In order to enhance the accuracy of the linear model, higher order timing models were introduced: [LLGP04] shows how to represent a circuit performance f of an analog or digital circuit by a transfer function of linear time invariant (LTI) system. The cumulative density function (cdf) and probability density function (pdf) can then be approximated by the step response and impulse response of the LTI system, respectively. In [CZNV05] an additional term is introduced in the canonical form, which captures nonlinear dependencies to process parameters. [ZCH$^+$05, ZCH$^+$06] offer a quadratic gate delay model and the according atomic operations *add* and *max*. A similar quadratic gate delay model is proposed [ZSL$^+$05] but in addition to that, also a quadratic wire delay model and a slightly different max operation is shown. Apart from these polynomial models, [CC05, CC07] proposes an analytical model for the gate delay and [BVGC06] enhances it by using Karhunen-Loève Expansion to model the spatial correlation between transistor parameters.

All methods mentioned above model only the delay and in some cases the slope. With downscaling feature sizes, however, this simplification of the waveforms leads to unacceptable inaccuracies of SSTA. Thus, similar as in the deterministic case, recent publications propose the application of current source models to SSTA: [FNP06] uses Markov chains to model the variations of the voltages at the output. The authors of

[ZXA$^+$07] model the variation of the waveform by basic operations like time shift, time scale, voltage shift and voltage scale. The parameters of these basic operations are then considered as random variables and propagated through the circuit in a block-based manner. In [GV08] a variational current source gate model is proposed but it remains unclear how the variation of the input signal propagates to the output.

Another wide field of research using the results of SSTA is the optimization of circuits by altering design parameters. As this is not the main focus of this work, only the work in [LLCP08] shall be mentioned. The authors describe a method of finding the arcs of the timing graph which are best suited to be optimized, i.e. up- or downscaled in order to improve the maximal frequency or the parametric yield.

1.2 Contributions of this Work

This work provides a novel path-based method for variational analysis of digital integrated circuits. The method works on a netlist of a particular path of a digital design, which includes all extracted parasitics after the layout. The intention is to build a highly accurate and reliable reference tool for SSTA. Therefore, higher runtimes are accepted, even though these runtimes inhibit the use as a tool for timing sign-off. The method extends an existing industrial tool, which provides the functionality of deterministic timing analysis, the partitioning of a path into separate stages and the determination of the critical path. This work adds the functionality of statistical analysis to this tool.

Different to existing methods, the entire waveform resulting from analog simulations as well as the complete interconnect structure and the dynamic load of the fanout are considered. The result is the information of the variation of the voltage at a number of points on the waveform at the output of a path. More precisely, the influence of each transistor parameter on the voltage at each specific point is available. This information can be used to compute the probability of violating the timing constraints or the optimization of the circuit to change the influence of specific parameters. The results show that the accuracy of the proposed method is comparable to analog Monte Carlo simulation but with a considerably shorter execution time. The method was implemented as an in-house reference tool at a leading manufacturer of integrated circuits with the need for evaluating available sign-off tools. The following publications resulted from this work: [SLS$^+$07, SKS08c, SKS08b, SKS08a]

Chapter 2

Problem Formulation

2.1 Delay

The main characteristic of digital circuits is the processing of two distinc signal states referred to as True/False, 1/0, On/Off or V_{dd}/V_{ss} depending on the abstraction level. The specification, which should be fulfilled by the circuit, can be represented by a logic function. The logic function consists of basic operations, e.g. NAND, NOR, etc., and each of these basic operations is represented by one logic gate in the circuit. Each gate has a specified number of inputs and most gates have one output. The gate library is the collection of all available gates and the optimal partitioning of the logic function to basic operations has to be found.

The gates are built of transistors which realize the basic functionality of switches. These switches connect or disconnect two circuit nodes and thereby control the flow of electrons, i.e. charge, depending on the signal value of the control pin, i.e. the gate, of the transistor. The transistor gates and various other parts of the physical implementation shown capacitive behavior. These capacitors have to be (dis-)charged whenever the signal value changes. This requires the transportation of charge through the transistors. The transported amount of charge per time and thus, the required time to (dis-)charge the capacitors depends on the physical properties of the transistors, e.g. threshold voltage or resistance in the on-state.

The delay of a logic gate is defined as the time from when the input signal crosses half the voltage swing, $(V_{dd} - V_{ss})/2$, to when the output signal crosses half the voltage swing. The gate delay depends on the physical properties of the transistors as well. Not only gates but also interconnects cause delay, which is defined similarly to the gate delay. The next section describes the problem of finding the delay of an entire circuit, given the delay of the individual logic gates and interconnects.

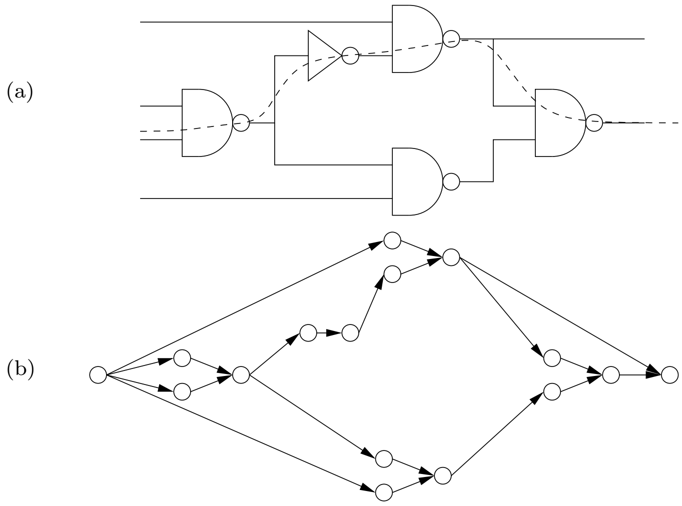

Figure 2.1: Example circuit (a) and corresponding timing graph (b)

2.2 Timing Analysis

The subject of timing analysis is to obtain the total delay of a circuit. An example circuit is shown in Figure 2.1(a). The circuit is represented by a timing graph which is a directed, acyclic graph (DAG) $\mathcal{G} = (N, E, n_s, n_f)$ as in Figure 2.1(b) [Hit82]. N denotes the set of nodes, E the set of edges, n_s the source node and n_f the sink node. The nodes represent primary input and output pins or the input and output pins of the gates. The edges represent the delay between the nodes, which can be either the gate delay or the interconnect delay.

The procedure of deterministic static timing analysis (STA) is to propagate the timing information through the timing graph in order to obtain the arrival time at the sink node n_f. Let p_i be a path in $\mathcal{G}$, represented by an ordered set of edges from the source node n_s to the sink node n_f. An example path is shown in a dashed line in Figure 2.1. The methodology proposed in this work is intended to serve as a verification tool to evaluate commercially available timing tools. Therefore, the analysis is confined to particular paths p_i. The path delay is not the delay of the entire circuit, but the result can be compared to the results of the path analysis of other tools. In addition, the proposed method enables the accurate analysis of special paths which are critical for the design.

In the following, the gate delay is merged with the interconnect delay, as it is analyzed in the same step and the two are physically closely related. The resulting timing graph for the example path from Figure 2.1 is shown in Figure 2.2.

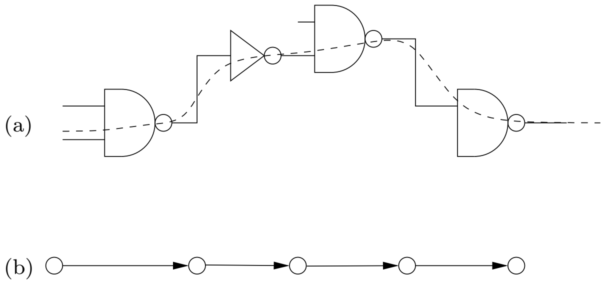

Figure 2.2: Example path (a) and corresponding timing graph (b)

The timing information is now propagated from the source node n_s to the sink node n_f through the path. Thus, the only operation which is needed is the transfer of timing information through one node. The operation for merging two or more incoming edges is not necessary in this path-based approach.

2.3 Variations

2.3.1 Uncertainties in Timing Analysis

During timing analysis of digital circuits, uncertainties arise from three different directions:

1. modeling and analysis errors: The modeling of devices and interconnects is based on simplifications, which cause inaccuracies. Also timing tools which employ these models possibly induce further errors.

2. variations in the manufacturing process: Parameters of devices and interconnects show die-to-die and within-die variations.

3. variations in the operating conditions: Supply voltage and temperature are examples of operating conditions which are subject to variations and influence the timing characteristics of the circuit.

The first influence occurs during the design process. During logic synthesis, buffer insertion, and place and route, timing analysis is performed. However, at each stage different inaccuracies are introduced due to effects such as undetected false paths, error in the cell delay, error in the extraction of interconnect parasitics, SPICE models, etc. All these effects cause the result of the timing analysis to be close to reality but not perfectly matched.

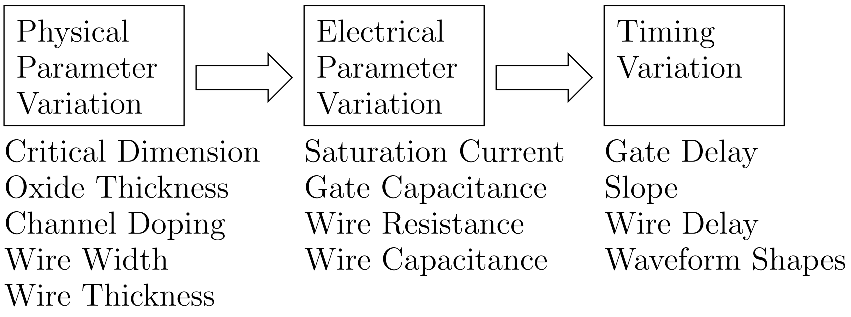

Figure 2.3: Variation of physical parameters causes variation in electrical parameters, which causes variation in timing parameters [BCSS08].

The next influence is due to process limitations or equipment imprecision. This influence is the main focus of SSTA and thus it is described in more detail in Section 2.3.2 below. Variations in operating conditions are especially relevant for circuits designed to operate in changing environments, e. g. mobile or automotive applications. Supply voltage and temperature influence the performance of a circuit and therefore, these influences have to be accounted for during the design. These variations in operating conditions are usually captured by corner-case margins, as each individual chip has to work properly under all temperature and supply voltage conditions.

2.3.2 Process Variations

With scaling feature sizes down to the nanometer region, it becomes harder to achieve accurately manufactured chips. Parameters that show variation caused by manufacturing inaccuracies can be classified into three categories:

1. physical parameters,

2. electrical parameters, and

3. performance or timing parameters.

The influence of physical parameters on electrical parameters and further on performance parameters is shown in Figure 2.3. The variation of physical parameters are due to various effects including chemical mechanical polishing (CMP). CMP is used to planarize insulating oxides and metal lines. Other effects are the optical proximity effect, which causes inaccuracies in structures which are smaller than the wavelength of light used in the lithography and lens imperfections in the optical system. Due to these and other effects, the parameters of devices and interconnects, such as gate length or critical dimension (CD), gate-oxide thickness, channel doping concentration, interconnect thickness and width, etc., show a substantial amount of variation.

The electrical characteristics of a device such as driving current or input capacitance depend on the physical parameters of the device. Thus, the variations of the physical parameters causes a variation of the electrical parameters. And, further, performance or timing parameters, e.g. delay, slope, or more complex quantities, in turn depend on the electrical parameters and the timing or performance parameters show variation as well caused by the variation of the electrical parameters.

It is important to note that more than one electrical parameter depends on a single physical parameter. For example, the wire resistance and capacitance both depend on the wire width and are thus correlated. The worst-case for wire resistance would mean a thin wire while the worst-case for the wire capacitance would mean a thick wire. Thus the worst-case for the wire delay – worst-case resistance and capacitance – is physically impossible. This shows how important it is to correctly consider the correlation between the electrical parameters.

Besides the three classes of parameters described above, a fourth class of parameters is introduced as *model* or *transistor parameters*. These are the parameters of the device models, e.g. BSIM4.3. They can be either physical parameters like channel length offset *xl*, electrical parameters like threshold voltage *vth0* or parameters which are derived from physical and electrical parameters, e.g. electrical gate equivalent oxide thickness *toxe*. Most tools operate on these model parameters and therefore, in the following mainly model parameters are used.

2.3.3 Physical Parameter Variation

The classification of physical parameter variation can be based firstly, on whether the variation is systematic or random or secondly based on the spatial range, i.e. long-distance or short-distance effects. The classification is shown in Figure 2.4.

1. *Systematic variations* are deterministic variations caused by well-understood physical effects. Systematic variations of physical parameters are mainly due to optical proximity effects, CMP and the associated metal fill. Since these effect are layout-dependent, they can be considered once the layout is finished. Thus, the consideration of systematic variations is possible at the end of the design process especially for timing sign-off. However, there is a need to include systematic effects into the analysis of earlier stages of the design when the

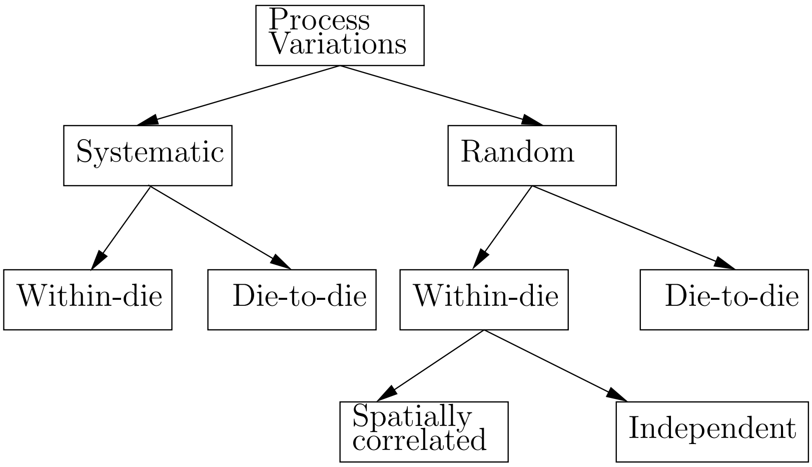

Figure 2.4: Classification of physical parameter variations [BCSS08].

layout is not available. Therefore, systematic variations are also considered statistically even though they occur due to deterministic effects.

2. *Random variations* are purely random variations of the process parameters. Before manufacturing, only the statistical characteristics are known and therefore, these variations have to be modeled by random variables. Examples for random variations are line-edge roughness (LER) and random dopant fluctuations (RDF).

Besides the classification into systematic and random variations, the variations can be classified according to the spatial characteristics.

1. *Die-to-die variations* (global, inter-die variations) affect each device on one die equally. This can be caused by a non-uniform etch concentration, which causes CD variation. The concentration might be different in the center of the wafer than on the edge causing a gradient in the CD. This gradient doesn't cause significant variation on one die but variation from one die located in the center of the wafer to another die which is located on the edge of the wafer. Other sources can be misalignment of the stepper carrying the lithography masks or a drift of the intensity of the laser. The spatial range of die-to-die variations could be further expanded to wafer-to-wafer, lot-to-lot, and fab-to-fab variation.

2. *Within-die variations* (local, intra-die variations) influence each device on the die differently. The CD of one transistor might be smaller than the CD of the transistor in close proximity to the first one.

Within-die random variations can be further classified into correlated and independent parameter variations. Within-die systematic variations can not be independent as they would be classified as random variations in that case. The reasons and implications of correlations are explained in the following section.

2.3.4 Statistical Parameters

Distribution of a Random Variable

This section describes the mathematical basis for the description of statistical parameters by random variables. The notations are derived from [HKO01] and more detailed information can be found in [Pap91]. Each statistical parameter s can be described by a random variable and is characterized by its *cumulative density function (cdf)* $F_s(s)$. The cdf shows the probability that the value of this parameter is below a certain threshold s_0:

$$F_s(s_0) = P(s \leq s_0) \tag{2.1}$$

For continuous random variables the cdf is a non-negative, nondecreasing continuous function. The values of F_s lie in the interval $0 \leq F_s(s) \leq 1$ and the limits are $F_s(-\infty) = 0$ and $F_s(+\infty) = 1$. Another way of describing the distribution is the *probability density function (pdf)* of a parameter s. The pdf is obtained as the derivative of the cdf:

$$p_s(s_0) = \left.\frac{dF_s(s)}{ds}\right|_{s=s_0} \tag{2.2}$$

The pdf can also be computed from a known pdf using the inverse relationship

$$F_s(s_0) = \int_{-\infty}^{s_0} p_s(\xi)\, d\xi \tag{2.3}$$

For simplicity $F(s)$ is used instead of $F_s(s)$ and $p(s)$ instead of $p_s(s)$ when no confusion may arise.
Assume now that $\mathbf{s} = (s_1, s_2, \ldots, s_n)$ is a random vector with the continuous random variables $s_1, s_2, \ldots, s_n$. The generalized form of the cdf becomes

$$F_\mathbf{s}(\mathbf{s_0}) = P(\mathbf{s} \leq \mathbf{s_0}) \tag{2.4}$$

and the pdf can be obtained by

$$p_\mathbf{s}(\mathbf{s_0}) = \left.\frac{\partial}{ds_1}\frac{\partial}{ds_2}\cdots\frac{\partial}{ds_n}F_\mathbf{s}(\mathbf{s})\right|_{\mathbf{s}=\mathbf{s_0}} \tag{2.5}$$

The distribution of two different random variables s_1 and s_2 can be described by the *joint cumulative distribution function (jcdf)*

$$F_{s_1,s_2}(s_{1_0}, s_{2_0}) = P(s_1 \leq s_{1_0}, s_2 \leq s_{2_0}) \tag{2.6}$$

Similarly to Equation 2.2, the *joint probability density function (jpdf)* can be defined by differentiating the jcdf F_{s_1,s_2} with respect to both variables s_1 and s_2. Thus, Equation 2.3 can be written as

$$F_{s_1,s_2}(s_{1_0}, s_{2_0}) = \int_{-\infty}^{s_{1_0}} \int_{-\infty}^{s_{2_0}} p_{s_1,s_2}(\xi, \eta)\, d\eta\, d\xi \tag{2.7}$$

Integration of the jpdf $p_{s_1,s_2}(s_1, s_2)$ over the variable s_2 leads to the *marginal density* $p_{s_1}(s_1)$ of s_1 and vice versa for $p_{s_2}(s_2)$:

$$p_{s_1}(s_1) \quad = \quad \int_{-\infty}^{\infty} p_{s_1,s_2}(s_1, \eta)\, d\eta \tag{2.8}$$

$$p_{s_2}(s_2) \quad = \quad \int_{-\infty}^{\infty} p_{s_1,s_2}(\xi, s_2)\, d\xi \tag{2.9}$$

Expectations and Moments

In most cases, the exact pdf of a model parameter s is unknown. The only data available are measurements from the production process. With this data expectations and higher order moments can be estimated. The expectation of a random variable s is defined by

$$E\{g(s)\} = \int_{-\infty}^{\infty} g(s)\, p_s(s) ds \tag{2.10}$$

The expectation can be estimated from measurements $g(s_j)$ by

$$E\{g(s)\} \approx \frac{1}{K} \sum_{j=1}^{K} g(s_j) \tag{2.11}$$

In the case of $g(s) = s$ Equation (2.10) describes the first moment or mean of the random variable

$$\mu_s = E\{s\} = \int_{-\infty}^{\infty} s\, p_s(s)\, ds \tag{2.12}$$

and the estimated mean value can be obtained by

$$\hat{\mu}_s = \frac{1}{K} \sum_{j=1}^{K} s_j \tag{2.13}$$

Higher order moments can be centered by subtracting the mean from the random variable. The centralized second moment is called variance and is defined by

$$\sigma_s^2 = E\{(s - \mu_s)^2\} \tag{2.14}$$

Considering more than one variable, vector notation for $\mathbf{s}$ is more convenient. The second moment of pairs of elements in $\mathbf{s}$ is called *covariance* and given by

$$c_{s_i,s_j} = E\{(s_i - \mu_{s_i})(s_j - \mu_{s_j})\} \tag{2.15}$$

When no confusion can arise, the notation c_{s_i,s_j} may be simplified to $c_{i,j}$ and for other variables respectively. The covariances of all statistical parameters s_i are put together to form the covariance matrix

$$\mathbf{C_s} = E\{(\mathbf{s} - \mu_\mathbf{s})(\mathbf{s} - \mu_\mathbf{s})^T\} \tag{2.16}$$

The elements of the covariance matrix normalized by their variances are called *correlation coefficients*

$$\rho_{i,j} = \frac{E\{(s_i - \mu_{s_i})(s_j - \mu_{s_j})\}}{\sigma_i \sigma_j} \tag{2.17}$$

Combining the correlation coefficients into a matrix yields the *correlation matrix* $\mathbf{R_s}$. The correlation matrix is symmetric ($\mathbf{R_s} = \mathbf{R_s}^T$) and positive semidefinite ($\mathbf{a}^T\mathbf{R_s}\mathbf{a} \leq 0$). Further, the eigenvalues of $\mathbf{R_s}$ are real and non-negative, all eigenvectors of $\mathbf{R_s}$ are real, and it is always possible to find mutually orthonormal eigenvectors.

Uncorrelatedness and Independence

The elements of a random vector are uncorrelated if the following condition holds for the covariance matrix $\mathbf{C_s}$:

$$\mathbf{C_s} = E\{(\mathbf{s} - \mu_\mathbf{s})(\mathbf{s} - \mu_\mathbf{s})^T\} = \mathbf{D} \tag{2.18}$$

with the diagonal matrix

$$\mathbf{D} = \mathrm{diag}(c_{1,1}, c_{2,2}, \ldots, c_{n,n}) = \mathrm{diag}(\sigma^2_{s_1}, \sigma^2_{s_2}, \ldots, \sigma^2_{s_n}) \tag{2.19}$$

For the special case of two random variables s_1 and s_2, the condition leads to the fact that their covariance $c_{1,2}$ is zero

$$c_{1,2} = E\{(s_1 - \mu_{s_1})(s_1 - \mu_{s_2})\} = 0 \tag{2.20}$$

or equivalently

$$E\{s_1 s_2\} = E\{s_1\}E\{s_2\} \tag{2.21}$$

It is important to note that the correlation coefficient does not capture the complete relationship between two random variables. In order to completely describe the relationship or dependence copulas are needed. Only for special cases like the Gaussian distribution, the correlation coefficient describes the entire dependence between two random variables.

Figure 2.5 shows distributions of two random variables p_1 and p_2. a) and b) show two Gaussian random variables which are a) uncorrelated and b) correlated. Note that the correlated variables are closer to a line through the center. c) and d) show two uniformly distributed random variables. In c) they are perfectly correlated and in d) the two variables are uncorrelated but not independent. These two random variables in d) are obtained by rotating two independent random variables s_1 and s_2:

$$\begin{bmatrix} p_1 \\ p_2 \end{bmatrix} = \begin{bmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{bmatrix} \begin{bmatrix} s_1 \\ s_2 \end{bmatrix} \tag{2.22}$$

with an arbitrary rotating angle α.

In general, uncorrelatedness does not imply independence of two random variables [Pap91]. Two random variables s_1 and s_2 are independent if and only if

$$p_{s_1,s_2}(s_1, s_2) = p_{s_1}(s_1)p_{s_2}(s_2) \tag{2.23}$$

The jpdf $p_{s_1,s_2}(s_1, s_2)$ must factorize into the product of their marginal densities $p_{s_1}(s_1)$ and $p_{s_2}(s_2)$. Two independent random variables satisfy the following condition:

$$E\{g(s_1)h(s_2)\} = E\{g(s_1)\}E\{h(s_2)\} \tag{2.24}$$

Comparing to (2.21) it can be seen, that independence is much stronger than uncorrelatedness and that uncorrelatedness only considers the second moment while independence considers all moments. This aspect becomes especially interesting when dealing with Gaussian random variables which are introduced in the following.

Gaussian Random Variables

In this work, it is assumed that model parameters can be modeled by Gaussian random variables with sufficient accuracy. Therefore, some details of the Gaussian distribution are highlighted in the following. The n-dimensional vector $\mathbf{s}$ of random variables is said to be Gaussian if the pdf of $\mathbf{s}$ has the form

$$p_{\mathbf{s}}(\mathbf{s}) = \frac{1}{(2\pi)^{n/2}(\det \mathbf{C_s})^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{s} - \mu_{\mathbf{s}})^T \mathbf{C_s}^{-1}(\mathbf{s} - \mu_{\mathbf{s}})\right) \tag{2.25}$$

For just one single variable s this simplifies to

$$p_s(s) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{s - \mu_s}{\sigma}\right)^2\right) \tag{2.26}$$

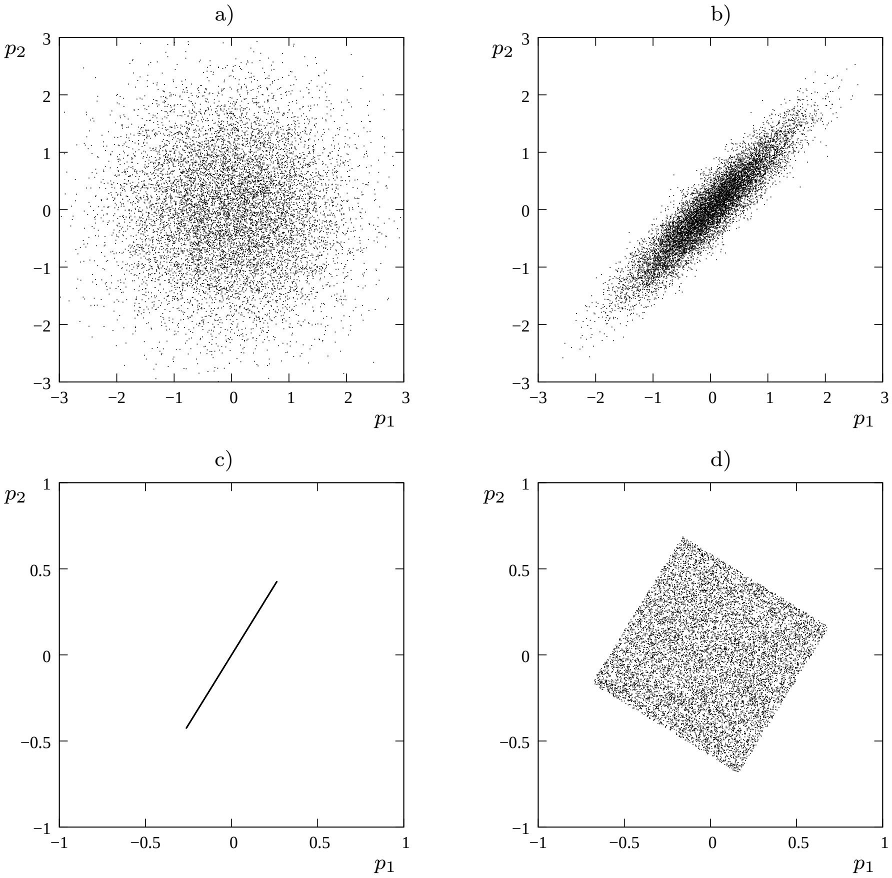

Figure 2.5: a) uncorrelated Gaussian variables b) correlated Gaussian variables c) perfectly correlated uniformly distributed variables d) uncorrelated but not independent uniformly distributed variables (see Equation (2.22))

Gaussian random variables have some special properties which make the assumption of Gaussianity attractive:

1. The distribution is completely described by only first and second moments. If the mean value μ_s and the standard deviation σ_s is known, the entire distribution can be obtained. This becomes an advantage in SSTA if it is assumed that signal arrival times are Gaussian as well. In which case, only the values for μ and σ of the arrival time have to be propagated through the circuit in order to compute a distribution of the arrival time at the output.

2. Gaussianity is preserved under linear operations. If $\mathbf{x}$ is a Gaussian random vector and $\mathbf{y} = \mathbf{A}\,\mathbf{x}$ a linear transformation, then $\mathbf{y}$ is also Gaussian with mean vector $\mu_{\mathbf{y}} = \mathbf{A}\,\mu_{\mathbf{x}}$ and covariance matrix $\mathbf{C}_{\mathbf{y}} = \mathbf{A}\,\mathbf{C}_{\mathbf{x}}\,\mathbf{A}^T$. This property is useful if linear sensitivities of timing quantities to transistor parameters are sufficient to achieve an acceptable accuracy. The transistor parameters can be modeled as Gaussian random variables and the timing characteristics like arrival time, slope or others will also remain Gaussian, which results in the advantages of the first property.

3. The third property regards the uncorrelatedness and geometrical structure of the Gaussian distribution. As mentioned earlier, the first and second moment are sufficient to describe the Gaussian distribution. If two Gaussian variables are uncorrelated according to Equation (2.21), Equation (2.23) holds as well and thus, both variables are independent.
   Any non-diagonal covariance matrix $\mathbf{C}_{\mathbf{s}}$, i.e. correlated random variables, can always be written in the form

   $$\mathbf{C}_{\mathbf{s}} = \mathbf{E}\,\mathbf{D}\,\mathbf{E}^T = \sum_{i=1}^{n} \lambda\,\mathbf{e}_i\,\mathbf{e}_i^T \tag{2.27}$$

   where $\mathbf{E}$ is an orthogonal matrix with columns $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_s$ being the n eigenvectors of $\mathbf{C}_{\mathbf{s}}$ and $\mathbf{D} = \mathrm{diag}(\lambda_1, \lambda_1, \ldots, \lambda_n)$ being the diagonal matrix comprised of the eigenvalues λ_i of $\mathbf{C}_{\mathbf{s}}$. Applying the rotation $\mathbf{E}$ to the centered vector s

   $$\mathbf{u} = \mathbf{E}^T(\mathbf{s} - \mu_{\mathbf{s}}) \tag{2.28}$$

   then yields a vector $\mathbf{u}$ of uncorrelated and hence independent Gaussian random variables.

The eigenvectors $\mathbf{e}_i$ and eigenvalues λ_i of $\mathbf{C}_{\mathbf{s}}$ can be interpreted geometrically as in Figure 2.6. All points on the jpdf with constant value ($p_{\mathbf{s}}(\mathbf{s})$ = const)

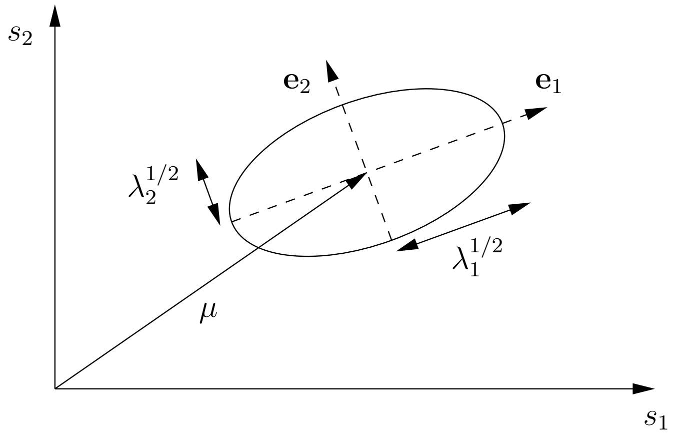

Figure 2.6: Geometric illustration of a two dimensional Gaussian distribution [HKO01].

> form hyperellipsoids centered at the mean value $\mu_{\mathbf{s}}$. The principal axes are parallel to $\mathbf{e_i}$ have the length λ_i. The figure intuitively shows, how a rotation decorrelates the random variables.

Using Gaussian random variables leads to sufficient accuracy of the proposed algorithm and also to a highly reduced complexity. Therefore, in the following Gaussianity is assumed and correlation is used to describe statistical dependence between random variables.

Correlations in SSTA

In the following, the implications of correlation for SSTA are highlighted. Correlations can occur between transistor parameters and between timing parameters such as arrival times and also between transistor and timing parameters. Correlations between arrival times can be caused by two effects:

1. *Spatial correlation* denotes correlation of transistor parameters depending on the distance to each other. Especially when systematic variations are modeled as random variables, they often show significant correlation. E.g. a gradient on the wafer leads to similar parameters for neighboring transistors. As the arrival times at the output of the cells are determined by the transistor parameters, these arrival times are also correlated.

2. *Topological correlations* arise when the circuit is analyzed in a block-based manner. If the output of a gate is connected to more than one following gate (fanout), the arrival times of all signals in the fanout cone are dependent on the

first fanout gate. Thus, all of these signals are correlated. The computation of the maximum arrival time at a reconvergence point becomes cumbersome.

As the proposed method is path-based, the problem of reconverging paths does not occur. Therefore, the following focuses on spatial correlations. If two parameters s_i and s_j of devices on one die belong to the class of die-to-die variations, they vary equally or in other words, these variables are perfectly correlated ($\rho_{i,j} = 1$). Within-die variations, on the other hand, can be either correlated ($0 < \rho_{i,j} < 1$) or completely independent ($\rho_{i,j} = 0$) as outlined in Section 2.3.3.

In order to capture correlations between parameters, the correlated parameters are represented by weighted sums of independent random variables. Two perfectly correlated parameters are described by one single random variable. If two parameters show arbitrary correlation, two independent random variables are contained in both weighted sums and the weights determine the correlation coefficient between the two parameters. To adapt this to spatially correlated transistor parameters, several methods have been published, which are all compatible with the SSTA method proposed in this paper. The method based on principle component analysis (PCA) and the method based on quad-tree partitioning of the chip will be explained in the following.

If the covariance matrix of the parameters is known and it is assumed that the random variables are Gaussian, PCA can be applied to generate the factors of the weighted sums to represent each correlated parameter as a sum of independent random variables [CS03b]. Let $\mathbf{z} \sim N(\mathbf{0}, \mathbf{C_z})$ be the vector of correlated Gaussian, zero mean random variables that have to be represented as a weighted sum of the independent, Gaussian, zero mean random variables $\mathbf{p} \sim N(\mathbf{0}, \mathbf{I})$ where $\mathbf{C_z}$ is the symmetric and positive definite covariance matrix obtained by an appropriate correlation model. $\mathbf{C_z}$ can be factored by the Cholesky factorization as

$$\mathbf{C_z} = \mathbf{K}\,\mathbf{K}^T \tag{2.29}$$

For more detailed information refer to [GvL85]. Using the Cholesky root $\mathbf{K}$ the vector $\mathbf{z}$ can be represented by

$$\mathbf{z} = \mathbf{K}\,\mathbf{p} \tag{2.30}$$

and $\mathbf{z}$ will also be Gaussian with mean

$$E\{\mathbf{z}\} = \mathbf{K}\,E\{\mathbf{p}\} = \mathbf{0} \tag{2.31}$$

and covariance matrix from (2.16):

$$E\{(\mathbf{Kp})(\mathbf{Kp})^T\} = E\{\mathbf{Kpp}^T\mathbf{K}^T\} = \mathbf{K}E\{\mathbf{pp}^T\}\mathbf{K}^T = \mathbf{KK}^T = \mathbf{C_z} \tag{2.32}$$

During the analysis the independent random variables **p** are used and after performing the SSTA, the parameter set **p** can be transformed back to the original set **z**.

A second option was introduced on page 17 and Figure 1.10 as a quad-tree model [ABZ03] where several layers are used to partition the chip into rectangles with lower layers showing a finer granularity of the partitioning. One independent random variable is assigned to each rectangle. Each parameter of a device on the chip is represented by a sum of the random variables which are assigned to the rectangles in which this particular device is located.

Canonical Sum

The result of these and similar methods is a sum of independent random variables and thus the processing of correlations can be considered a preprocessing step before independent random variables are assumed for timing analysis. As it simplifies the following, zero mean random variables are used and the mean is considered separately from the sum.
This sum of independent random variables is called *canonical sum* as in [ABZ03, VRK$^+$04] and can be written for parameter p_i as:

$$p_i = p_{i,0} + \delta p_i = p_{i,0} + \sum_{j=0}^{N} \alpha_{i,j} \delta p'_j \tag{2.33}$$

Wherein $\delta p'_j$ are the independent, zero mean random variables, resulting from the correlation model above and $p_{i,0}$ is the mean of p_i. A more compact writing can be obtained using vector notation:

$$\mathbf{p} = \mathbf{p_0} + \delta \mathbf{p} \tag{2.34}$$

Using this representation of varying transistor parameters, the next step is the mapping to performance or timing parameters u_i. A linear model is chosen to represent this mapping. Such a linear model is valid only if the variations are not too large. Experiments on industrial process data showed that variations stay well within this linear region. If the transistor parameters are modeled by a canonical sum and the mapping to performance parameters is linear, the performance parameters may as well be represented by a canonical sum:

$$u_i = u_{i,0} + \delta u_i = u_{i,0} + \sum_{j=0}^{N} \alpha_{i,j} \delta p'_j \tag{2.35}$$

The parameters $\delta p'_j$ denote the independent parameters from the correlation model or transformation of random variables. In section 3.4.1 it is shown how the canonical sums of the performance parameters u_i can be transformed into an expanded representation leading to a smaller variation of these parameters.

The SSTA problem is now to obtain the coefficients $\alpha_{i,j}$ for each cell and traverse the circuit in order to obtain the $\alpha_{i,j}$ for the desired performance parameters u_i at the output of the circuit.

2.4 Waveform Relevance and Suitable Modeling

Presenting the state-of-the-art in Section 1.1, the chronological development from delay to slope and further to the consideration of waveform shapes was sketched. In the problem formulation so far, only timing information or performance parameters were mentioned. This should now be specified by answering the following question: How accurately should the signals in the circuit be modeled?

Traditional approaches use arrival time and slope but the influence of wire resistance is increasing and thus the relevance of waveforms. In order to quantify the influence of signal shapes, a NAND gate with two inputs from an industrial library is simulated with two differently shaped input waveforms. To ensure that these waveforms are realistic, they are obtained by simulating the same gate twice with different values of the resistance in the RC-load. The two resulting waveforms were shifted and scaled such that the arrival time and slope matched. Figure 2.7 shows these two input waveforms with equal arrival time and slope as well as the resulting output waveforms. The difference of the delay for the two different waveforms at the input is approximately 10%. This comparison shows that arrival time and slope are not sufficient to model the waveforms when increasing wire resistance causes deviation from standard waveform shapes. The problem of waveform based SSTA can be divided into three parts similar to the classification mentioned earlier:

1. A suitable *waveform model* has to be found which can describe more details of the waveform. The sensitivity coefficients in Equation (2.35) have to be computed and propagated through the circuit. The result is the analysis of the waveform variation of the output of the circuit.

2. The *driver model* has to be extended as well because classical tables can not deal with waveform shapes. More sophisticated methods such as current source models suffer from simplifications that cannot guarantee high accuracy for all possible application scenarios.

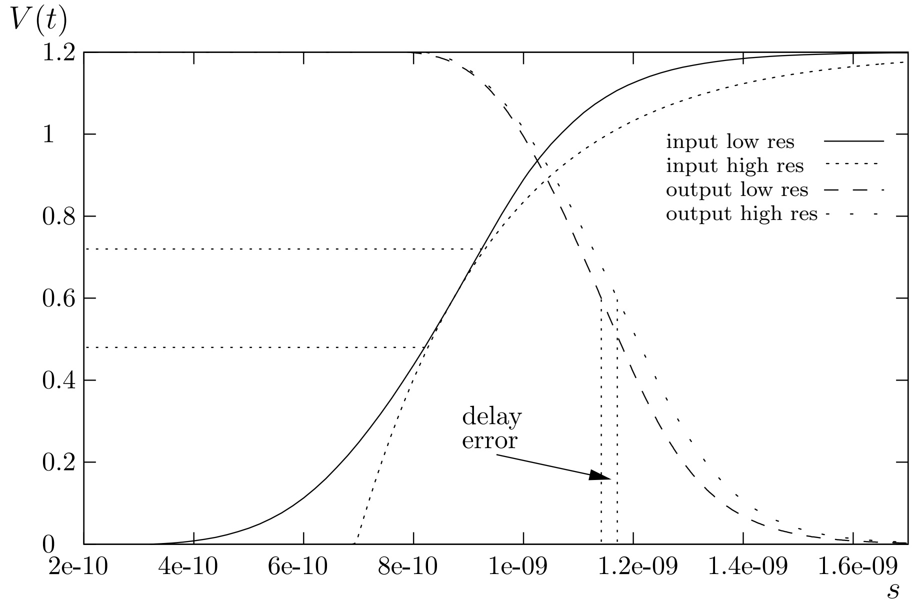

Figure 2.7: Two waveforms with equal arrival time and slope obtained by altering the resistance of the interconnect

3. The *receiver model* has to be refined. As shown earlier, a simple capacitor can not represent the resistance of the wires and also nonlinear input capacitances of gates cause severe problems when using approximated models.

These three problems will be addressed in the next section. The choice how to solve each problem is taken in favor of accuracy rather that runtime. The reason for that is that a reference tool is needed to evaluate commercially available tools in cases where Monte Carlo analysis is not possible due to runtime restrictions.

2.5 Integration into Industrial Context

The aim of this work is to provide a methodology which is not only of academic interest but is also of use to industry partners. This methodology is intended to extend current industry tools. Such a tool is the *Path Delay Calculator (PDC)* of one of the leading global semiconductor manufacturers. The PDC is a block-based STA reference tool based on transistor level simulations. It considers waveforms and complex driver and receiver models including entire interconnect structures. In this work the PDC is extended by statistical analysis to a path-based SSTA reference tool. Future projects could endeavor to adopt the block-based procedure, which is not the scope of this work.

Chapter 3

Waveform Based Timing Analysis

This chapter describes the core ideas of a novel approach for path-based Statistical Static Timing Analysis. The different aspects of modeling are highlighted followed by the description of the nominal STA. After that, the consideration of parameter variations is introduced and how these variations are propagated through a path. At the end of the path the results are post-processed in order to compare them in a quantitative manner to the results of a Monte Carlo analysis.

3.1 Modeling

As mentioned earlier, three models are needed for SSTA: Waveform model, driver model and receiver model. These models are needed to traverse through a circuit in a path-based manner. In a standard STA the timing properties of the individual cells are obtained by analog simulations during library characterization. The current industry standard uses two parameters to model the timing of the cells: The cell behavior is parametrized by the delay and slew of the output signal as a function of the slew of the input signal and the capacitive output load. The characterization data is stored in look-up tables for each signal path through the cell and for each standard cell available in the library. This data is then retrieved from the tables by the STA tool. In contrast to this approach, which is of limited accuracy due to its simplified waveform, receiver load, and interconnect modeling, a different, more accurate approach is followed here. It is based on successive analog simulations along the given timing paths. The circuit is traversed along the path gate by gate. For each gate a transistor level simulation is performed to obtain the exact nominal output waveform $V_{\mathrm{nom}}^{\mathrm{out}}(t)$, i. e. nominal output voltage. For the propagation of statistical variations, three basic steps can be identified: Determining the influence of the parameter variations of the current gate on the output of this gate, determining the influence of the input waveform variation of the current gate on the output

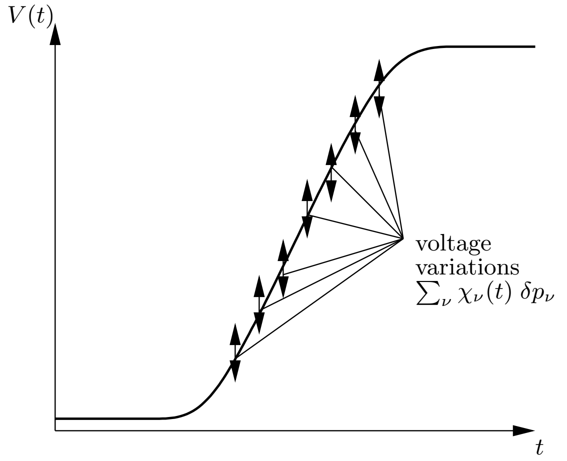

Figure 3.1: Waveform variations modeled by canonical sums of transistor parameters

of this gate, and the addition of these two influences with correct consideration of correlations. This yields the variation of the output signal which can be used as the variation of the input signal of the next gate. In the following, the three necessary parts of modeling are described in more detail and after that, the traversal through a path is explained.

3.1.1 Waveform Model

This work endeavors to provide a reference tool, which achieves the highest possible accuracy but with a considerably shorter execution time than Monte Carlo SPICE simulation. Thus, the amount of simplification is reduced to a minimum. The tool is based on an analog simulator and thus, the best choice for the waveform model is to consider the entire result of the analog simulator. The PDC follows this concept and saves the entire analog waveform and uses it in the further analysis without any simplifications. In order to model the variations of this waveform, specific points on the waveform are selected and the variation at these points caused by the variations of transistor parameters is modeled by variations of voltage at these points of time on the waveform. The voltage variations are represented by canonical sums. This waveform model is depicted in Figure 3.1. By changing the number of selected points in time the accuracy can be traded in for runtime.

3.1.2 Driver Model

As for the waveform model, the focus is to achieve highest accuracy possible. Thus, the driving gates are not modeled by look-up tables or current source models. The entire transistor level netlist is considered. This model includes all parasitics in the gates as well as non-linearities and omits common sources of modeling inaccuracies. The analog waveform is used as input of these gates and the output is also an analog waveform.

3.1.3 Receiver Model

All relevant effects should be covered by the receiver model. State-of-the-art tools model the receiver as an effective capacitance, e. g. models based on look-up tables, or πRC or similar structures, e. g. current source models. Effects like resistive shielding of interconnects and non-linear capacitances at the input of logic cells due to the Miller effect have to be considered by these simple models, which leads to inevitable inaccuracies in the receiver modeling. To overcome this problem, the entire transistor level netlist of the receiving structure is used. This includes all cells connected to the output of the current cell as well as the entire interconnect structure with all parasitics extracted from the layout.

3.2 Waveform-Based Nominal STA

As mentioned above, the proposed SSTA methodology is based on analog simulations in order to achieve the most accurate result possible. The variational analysis is based on the nominal waveform. This can be also seen as the operation point. All sensitivities and other computations have to be executed assuming that particular point. As the variation of the waveform is considered, the nominal waveform denotes the operation point or the offset for the zero mean random variables. Therefore, the first step is to determine the nominal waveform for the selected path. To take into account the fact that the waveform depends on the dynamic load of the respective gate, all simulations are performed on *stages* which include the cell under consideration plus all relevant elements and interconnect structures in the fanout as can be seen in Figure 3.2.

The first stage is simulated by the analog simulator using a given input waveform. The result is the nominal waveform at the output of the first stage $V_{\mathrm{nom}}^{\mathrm{out}}(t)$. To analyze the timing of the path, the process is repeated: The output signal of the stage serves as input for the next stage and so forth. As each gate contains only a small number of transistors, the individual simulations can be completed with acceptable simulation times. As the complete waveforms as well as the full influence of the interconnects are taken into account, this path-based STA approach reaches accuracies that are comparable to that of SPICE simulations of the whole paths,

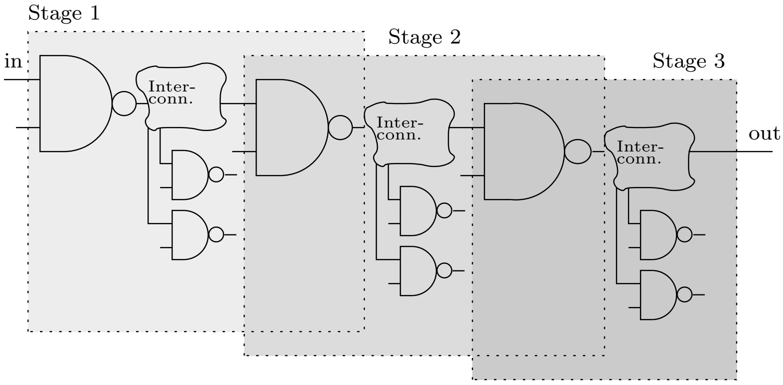

Figure 3.2: Partitioning of a path into stages

while being much faster due to its successive nature [BM06].

3.3 Parameter Variations

3.3.1 Linear Modeling of Variations

As outlined earlier, the transistor parameters are subject to statistical variations. Thus, the parameters $\mathbf{p} = (p_1, \ldots, p_N)$ of the transistors produced in silicon exhibit deviations $\delta\mathbf{p}$ from the nominal values $\mathbf{p_0}$:

$$\mathbf{p} = \mathbf{p_0} + \delta\mathbf{p} \tag{3.1}$$

As a consequence of these variations, the timing behavior of the cells changes as well. In Section 2.3.4 on page 35 the canonical sum is introduced to represent the linear model of performance parameters, which is state-of-the-art.

In the following a similar linear analysis concept is described but in contrast to existing methods the entire waveforms involved in the signal propagation are considered. Assume that a signal at a given node in the circuit is given by a voltage as a function of time and transistor parameters, $V = V(t, \mathbf{p})$. The canonical sum of performance parameters from Equation (2.35) can be rewritten using time continuous weights $\chi_\nu(t)$ to represent the variation of the nominal waveform:

$$V(t, \mathbf{p_0} + \delta\mathbf{p}) \approx V(t, \mathbf{p_0}) + \sum_{\nu=1}^{N} \chi_\nu(t)\ \delta p_\nu \tag{3.2}$$

with

$$\chi_\nu(t) \equiv \left.\frac{\partial V(t, \mathbf{p_0} + \delta\mathbf{p})}{\partial \delta p_\nu}\right|_{\delta\mathbf{p}=0} = \frac{\partial V(t, \mathbf{p})}{\partial p_\nu} \tag{3.3}$$

In the case of time-discrete values of the signal and the respective sensitivities, Equation (3.2) for time t_i becomes:

$$V_i = V(t_i, \mathbf{p_0} + \delta\mathbf{p}) \approx V(t_i, \mathbf{p_0}) + \sum_{\nu=1}^{N} k_{i,\nu}\, \delta p_\nu \tag{3.4}$$

wherein the $k_{i,\nu}$ are the weights of parameter δp_ν and time t_i. The voltage variation δV_i at time t_i is a zero mean random variable defined as

$$\delta V_i = \sum_{\nu=1}^{N} k_{i,\nu}\, \delta p_\nu \quad . \tag{3.5}$$

Using matrix notation, the vector of zero mean random variables $\delta\mathbf{V}$ denoting the voltage variations at different times can be written as

$$\delta\mathbf{V} = \mathbf{K}\,\delta\mathbf{p} \tag{3.6}$$

with the coefficients matrix $\mathbf{K}$.

3.3.2 Statistical Variations

In a statistical framework, the deviations $\delta\mathbf{p}$ are assumed to be random variables, which vary statistically from die to die, wafer to wafer, and lot to lot according to given statistics. The statistical variations are represented by canonical sums as introduced in Section 2.3.4. It is also shown in Section 2.3.4 how this representation allows the consideration of correlations.

The current standard of library characterization distinguishes between global and independent local variations. For this reason it is straightforward to divide the transistor parameters into global parameters and local parameters. The determination of correlation coefficients between parameters of different transistors is still a subject of current research and not integrated into the standard process of library characterization. Thus, the currently available data does not allow the consideration of correlation between parameters of different transistors. Note, however, that spatial correlation can be integrated into the proposed method easily once the data is available.

Global variables (denoted by z_ν, $\nu = 1, \ldots, M$) describe parameters which are constant within one die, but vary from die to die, wafer to wafer, and lot to lot. They are shared by all the cells in the design and induce correlations between the various cells of a given design since changes in these variables affect all these cells in a coordinated manner. Local variables (in the following denoted by $\zeta_\mu^{(i)}$, $\mu = 1, \ldots, L$), on the other hand, model independent variations from transistor to transistor within the same die (e.g. due to dopant fluctuations etc.). They generate variations of cell properties which are independent from cell to cell. Therefore, a general linear response expression for a quantity Q that is characteristic for the given cell i,

$$Q^{(i)}(\mathbf{p} + \delta\mathbf{p}) \approx Q_{\text{nom}}^{(i)} + \sum_{\nu=1}^{N} \alpha_\nu^{(i)} \delta p_\nu^{(i)} \tag{3.7}$$

more explicitly splits up into two types of contributions,

$$Q^{(i)}(\mathbf{p} + \delta\mathbf{p}) \approx Q_{\text{nom}}^{(i)} + \sum_{\nu=1}^{M} \alpha_{\nu,\text{global}}\, \delta z_\nu + \sum_{\mu=1}^{L} \alpha_{\mu,\text{local}}^{(i)}\, \delta\zeta_\mu^{(i)} \tag{3.8}$$

which correspond to the global and local variations respectively. This careful distinction between global and local parameters becomes important in Section 3.4.2 when canonical sums have to be added.

3.4 Path-based Analysis

This section provides a methodology for path-based statistical timing analysis. After the computation of the nominal waveform, which was described in Section 3.2, the statistical variations are considered in the following. First, the behavior of a single cell is explained. This procedure is successively applied on the cells of a particular path.

3.4.1 Single Stage Analysis

As mentioned in Section 3.1, three basic steps have to be executed for the analysis of a single stage:

1. computation of the variation of the output waveform due to variations of the transistor parameters $\delta\mathbf{p}^{\text{stage}}$ of the stage itself,
2. computation of the variation of the output waveform due to variations of the input waveform $\delta\mathbf{p}^{\text{input}}$, and
3. addition of the two influences above.

These three steps will be explained in the following. The different contributions are sketched in Figure 3.3. In the next section a method to model waveform variations of the output of a single stage by linear response methods will be derived.

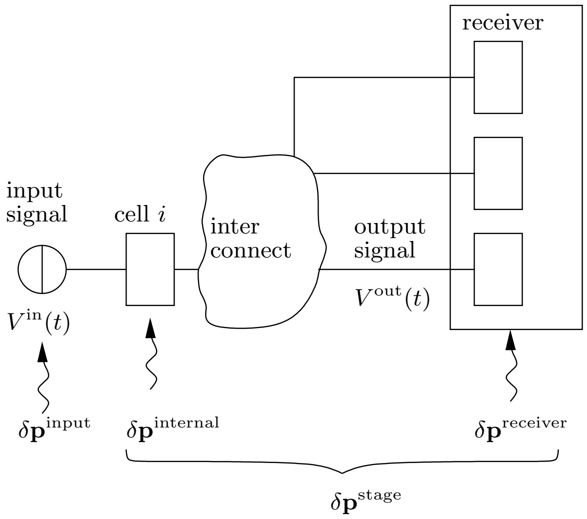
receiver
input signal
cell i
inter connect
output signal
$V^{\text{in}}(t)$
$V^{\text{out}}(t)$
$\delta\mathbf{p}^{\text{input}}$
$\delta\mathbf{p}^{\text{internal}}$
$\delta\mathbf{p}^{\text{receiver}}$
$\delta\mathbf{p}^{\text{stage}}$

Figure 3.3: Single stage with parameter variations

Stage Parameter Variations

Without any variations, i.e. $\delta\mathbf{p} \equiv \mathbf{0}$, the nominal input waveform $V^{\text{in}}_{\text{nom}}(t)$ generates the nominal output waveform $V^{\text{out}}_{\text{nom}}(t)$. Taking into account the variation of the transistor parameters of the stage leads to a variation from this nominal waveform. The following approach is based on an analog simulator which can compute the sensitivities $\partial V_i/\partial p_\nu$ of voltages $V_i = V(t_i)$ at the output at time t_i to transistor parameters p_ν. The chosen simulator employs adjoint networks in order to compute sensitivities of measurements to a large number of parameters with little overhead [DR69, PSD70, PRV95]. These sensitivities are the coefficients needed in order to formulate the canonical sum for the variation of each voltage δV_i at time t_i

$$\delta V_i = \delta V(t_i) = \sum_{\nu=1}^{N} \frac{\partial V_i}{\partial p_\nu}\delta p_\nu = \sum_{\nu=1}^{N} k_{i,\nu}\delta p_\nu \tag{3.9}$$

The voltage variations $(\delta V_1, \delta V_2, \ldots, \delta V_M)$ will be interpreted as a continuous function $\delta V(t)$ in order to obtain a more compact and general formulation:

$$\delta V(t) = \sum_{\nu=1}^{N} \frac{\partial V(t)}{\partial p_\nu}\delta p_\nu = \sum_{\nu=1}^{N} \chi_\nu(t)\delta p_\nu \tag{3.10}$$

wherein $\chi_\nu(t)$ is the time dependent sensitivity of the output voltage to the transistor parameter p_ν. This linear approximation is only valid as long as the parameter deviations are sufficiently small.

However, for longer paths this linearization error becomes worse. The reason is that the main effect of parameter variations is a time shift of the waveform. This

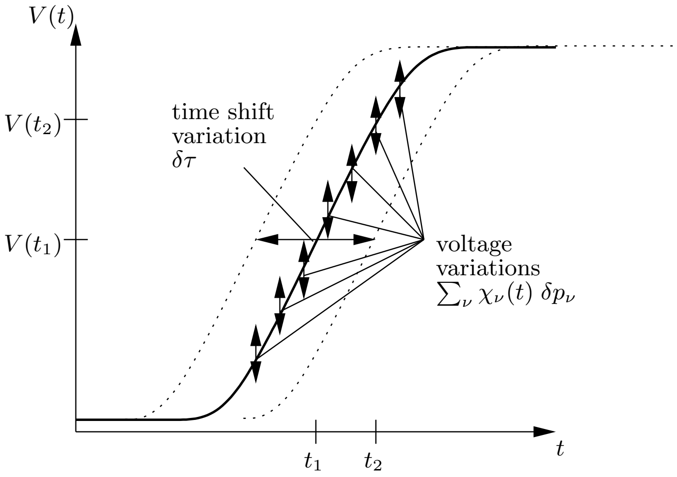

Figure 3.4: Example waveform with translation and voltage variations

time shift of the nominal waveform is a zero mean random variable as well. It causes all voltages to vary in the same direction. For longer paths this time shift variation becomes larger and thus the variation of the voltages due to the time shift becomes larger as well. The result is a large variation that does not yield much information and can be represented more efficiently. Such an approach is the separate consideration of the arrival time variation. Figure 3.4 shows such an extended representation with separate time shift variation and voltage variations.

The aim is to determine the variation in voltage caused by the time shift variation. In case of small variations, the time variation can be transformed into a voltage variation by multiplying with the derivative with respect to time of the waveform

$$\Delta V(t_i) \approx \left.\frac{\partial V(t)}{\partial t}\right|_{t=t_i} \Delta t_i \tag{3.11}$$

If a change in parameters $\mathbf{\Delta p}$ causes only a time shift $\Delta\tau$ and no deformation then holds

$$V(t, \mathbf{p} + \mathbf{\Delta p}) = V(t - \Delta\tau, \mathbf{p}). \tag{3.12}$$

The example in Figure 3.4 should be used to clarify this. At time t_2 the nominal voltage is $V(t_2)$. The aim is to determine the voltage after the waveform is shifted to the right by a positive value of $\Delta\tau = t_2 - t_1$ caused by a change of the parameters $\mathbf{\Delta p}$. Note that Δ denotes deterministic alteration and δ denotes the respective random variable. For the voltage at time t_2 follows:

$$V(t_2, \mathbf{p} + \mathbf{\Delta p}) = V(t_2 - \Delta\tau, \mathbf{p}) = V(t_1, \mathbf{p}) \tag{3.13}$$

Thus, a positive time shift $\Delta\tau$ results in a negative Δt. For a rising (falling) waveform this causes a negative (positive) voltage shift. This introduces a minus sign changing (3.11) to

$$\Delta V(t_i) \approx -\left.\frac{\partial V(t)}{\partial t}\right|_{t=t_i} \delta\tau \tag{3.14}$$

Computing the sensitivity of the arrival time to the transistor parameters $\partial\tau/\partial p_j$ as well, the influence of the transistor parameter variations on the time shift can be determined and a canonical sum for the variation of the time shift can be written as

$$\delta\tau = \sum_{\nu=1}^{N} \frac{\partial\tau}{\partial p_\nu}\delta p_\nu = \sum_{\nu=1}^{N} \vartheta_\nu \delta p_\nu \tag{3.15}$$

The time shift variation influences all points on the waveform equally. Thus, the voltage variation $\delta\tilde{V}(t)$ caused by the time shift variation $\delta\tau$ can be computed using Equation (3.11):

$$\delta\tilde{V}(t) = -\sum_{\nu=1}^{N} \frac{\partial V(t)}{\partial t}\frac{\partial\tau}{\partial p_\nu}\delta p_\nu = -\frac{\partial V(t)}{\partial t}\delta\tau \tag{3.16}$$

Knowing the voltage variation which is caused by the variation of the time shift, this contribution can be subtracted from the sensitivities $\chi_\nu(t)$ in Equation (3.10)

$$\chi'_\nu(t) = \chi_\nu(t) - \left(-\frac{\partial\tau}{\partial p_\nu}\frac{\partial V(t)}{\partial t}\right) \tag{3.17}$$

leading to a new expression with a contribution from the time shift variation and one from the reduced parameter variations.

$$\begin{aligned} \delta V(t) &= -\sum_{\nu=1}^{N} \frac{\partial V(t)}{\partial t}\frac{\partial\tau}{\partial p_\nu}\delta p_\nu \quad + \quad \sum_{\nu=1}^{N} \chi'_\nu(t)\delta p_\nu \\ &= -\frac{\partial V(t)}{\partial t}\delta\tau \quad + \quad \sum_{\nu=1}^{N} \chi'_\nu(t)\delta p_\nu \end{aligned} \tag{3.18}$$

The canonical sum for the voltage variations in the right hand side of Equation (3.18) now represents only the voltage variation without the variation caused by the time shift. The variation caused by the time shift is considered separately in the variable $\delta\tau$. This variable is computed additionally and propagated through the circuit as described in the following.

Input to Output Transfer of Variations

After the examination of the influence of transistor parameter variations of the current stage on the output waveform, this section deals with the influence of the input waveform variations on the variations of the output waveform. As different canonical sums are used for input and output variations, the sensitivities ϑ_ν and χ_ν are replaced by $\vartheta_\nu^{\text{in}}$, $\vartheta_\nu^{\text{stage}}$, $\vartheta_\nu^{\text{out}}$, χ_ν^{in}, χ_ν^{stage} and χ_ν^{out} for input and output sensitivities respectively. The variation of the arrival time at the input $\delta\tau^{\text{in}} = \sum_{\nu=1}^{N} \vartheta_\nu^{\text{in}} \delta p_\nu$ is simply added to the variation of the arrival time resulting from the variation of the stage parameters $\delta\tau^{\text{stage}} = \sum_{\nu=1}^{N} \vartheta_\nu^{\text{stage}} \delta p_\nu$.

$$\delta\tau^{\text{out}} = \sum_{\nu=1}^{N} \left(\vartheta_\nu^{\text{in}} + \vartheta_\nu^{\text{stage}}\right) \delta p_\nu \tag{3.19}$$

In the following the variation of the time shift is left out for the sake of clarity.

Assume that variations in the transistor parameters of previous stages lead to a change in the signal at the input of the current stage.

$$V^{\text{in}}(t) = V_{\text{nom}}^{\text{in}}(t) + \delta V^{\text{in}}(t) \tag{3.20}$$

As the parameters of the current stage are not considered here, the random vector $\delta\mathbf{p}$ denotes the variation of parameters from the previous stages. As long as the alterations δp_ν of the corresponding transistor parameters, which determine the input signal, are sufficiently small, the signal deformation $\delta V^{\text{in}}(t)$ depends linearly on these deviations. Assume that the signal deformation is given by a canonical expression of the form

$$\delta V^{\text{in}}(t) = \sum_{\nu=1}^{N} \chi_\nu^{\text{in}}(t) \;\; \delta p_\nu \tag{3.21}$$

where the sensitivity coefficients $\chi_\nu(t)$ are given by the computations of the previous stages. This variation of the input waveform causes a variation of the waveform at the output of the current stage:

$$V^{\text{out}}(t) = V_{\text{nom}}^{\text{out}}(t) + \delta V^{\text{out}}(t) = V_{\text{nom}}^{\text{out}}(t) + \sum_{\nu=1}^{N} \chi_\nu^{\text{out}}(t) \;\; \delta p_\nu \tag{3.22}$$

with

$$\chi_\nu^{\text{out}}(t) = \frac{\partial V_{\text{nom}}^{\text{out}}(t)}{\partial p_\nu} \tag{3.23}$$

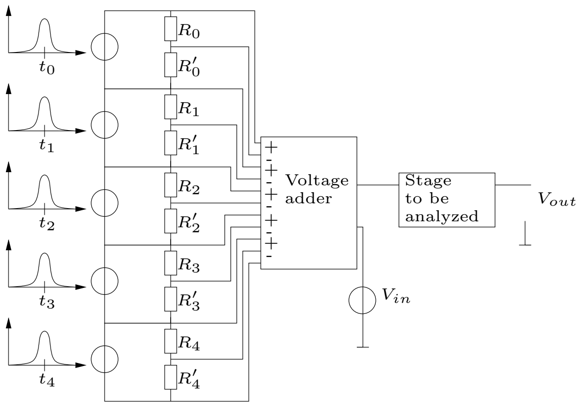

Figure 3.5: Auxiliary circuit to determine the sensitivities to input voltages.

The problem is to compute the sensitivity $\chi_\nu^{\mathrm{out}}(t)$ of the output voltage to the input parameter variation. This computation has to be efficient and has to seamlessly integrate into the already existing path delay calculator (PDC) as mentioned in Section 2.5. Thus, the aim is to find a suitable way of computing these sensitivities using the analog simulator. This can be achieved by formulating the needed sensitivities as sensitivities of the output voltage to device parameters. These sensitivities can be computed directly by the analog simulator. It is currently not possible to compute sensitivities of the output voltage to the input voltage using adjoint network analysis.

3.4.2 Sensitivities from SPICE Simulations

Two options can be identified to formulate sensitivities based on device parameters. The idea behind both options is to use voltage dividers with a voltage source for each voltage divider and to compute the sensitivities of the output voltage to one of the two resistors of each voltage divider. The resulting auxiliary circuit can be seen in Figure 3.5. With the correct scaling this directly yields the appropriate sensitivities.

The first option approximates the transfer characteristic of a cell by the time dependent linear transfer function and obtains the output variation by integrating over the transfer function and the input variation. The second option uses a more direct and smoother approach, which yields the influence of the input variations to the output voltage directly.

Approximating the Transfer Function

Each stage is considered as a linear system with input voltage variation $\delta V^{\mathrm{in}}(t)$ as input variable and output voltage variation $\delta V^{\mathrm{out}}(t)$ as the output variable. The system is described by the linear operator T [Unb72]:

$$\delta V^{\mathrm{out}}(t) = T[\delta V^{\mathrm{in}}(t)] \tag{3.24}$$

The system is assumed to be linear but is not time invariant because the input to output characteristics change while the voltages change during the signal transition. The idea is to estimate the impulse response $h(t, \tau)$ by measurements and to compute the output variation by integrating the product of input variation and impulse response. This will be derived in the following.

From the linearity of the operator T follows

$$T\left[\sum_{\nu=1}^{N} k_\nu \, \delta V^{\mathrm{in},(\nu)}(t)\right] = \sum_{\nu=1}^{N} k_\nu \, T\left[\delta V^{\mathrm{in},(\nu)}(t)\right] \tag{3.25}$$

Thus can be shown:

$$T\left[\int_a^b k(\nu)\, \delta V^{\mathrm{in}}(t,\nu)\, d\nu\right] = \int_a^b k(\nu)\, T\left[\delta V^{\mathrm{in}}(t,\nu)\right] d\nu \tag{3.26}$$

The impulse function $\delta(t)$ is defined by:

$$x(t) = \int_{-\infty}^{\infty} f(\tau)\, \delta(t-\tau) d\tau \tag{3.27}$$

The response of the system to the impulse function is

$$h(t,\tau) = T[\delta(t-\tau)] \tag{3.28}$$

The output signal $\delta V^{\mathrm{out}}(t)$ can be written as operator T applied to the input signal $\delta V^{\mathrm{in}}(t)$:

$$\delta V^{\mathrm{out}}(t) = T[\delta V^{\mathrm{in}}(t)] \tag{3.29}$$

With Equations (3.27) and (3.26) follows:

$$\begin{aligned} \delta V^{\mathrm{out}}(t) = T[\delta V^{\mathrm{in}}(t)] &= T\left[\int_{-\infty}^{\infty} \delta V^{\mathrm{in}}(\tau)\delta(t-\tau) d\tau\right] \\ &= \int_{-\infty}^{\infty} \delta V^{\mathrm{in}}(\tau) T\left[\delta(t-\tau)\right] d\tau \end{aligned} \tag{3.30}$$

and with Equation (3.28):

$$\delta V^{\mathrm{out}}(t) = \int_{-\infty}^{\infty} \delta V^{\mathrm{in}}(\tau) h(t,\tau) d\tau \tag{3.31}$$

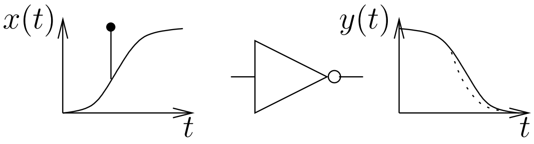

Figure 3.6: Impulse Response

As the system is causal, the upper limit can be changed to t:

$$\delta V^{\text{out}}(t) = \int_{-\infty}^{t} \delta V^{\text{in}}(\tau)h(t,\tau)d\tau \tag{3.32}$$

The following describes how to obtain the impulse response $h(t,\tau)$ needed for Equation (3.32). The impulse response can be obtained by exciting the stage with a single Dirac impulse at the input at time τ and measuring the voltage at the output at time t. As the impulse occurs during the change of the signal value, the impulse response $h(t,\tau)$ depends on τ. As the number of simulations should remain low, it is desired to obtain $h(t,\tau)$ for all τ in one simulation run. The idea is to use the built-in sensitivity analysis of the simulator and compute the sensitivity of the output to the impulse at the input. But as the simulator is not able to determine sensitivities with regard to input voltage, an analyzer circuit needs to be constructed which is shown in Figure 3.5.

Instead of continuous signals $\delta V^{\text{in}}(t)$ and $\delta V^{\text{out}}(t)$ only a number of samples $\delta V_j^{\text{in}} = \delta V^{\text{in}}(t_j)$ and $\delta V_i^{\text{out}} = \delta V^{\text{out}}(\tau_i)$ are considered. For each time τ_i there is one voltage divider with resistors R_i' and R_i and voltage sources $\delta(t-\tau_i)$. As a real Dirac impulse is numerically problematic, a Gaussian input signal $V_i(t)$ is used:

$$\delta(t-\tau_i) \approx V_i(t) = \frac{\theta}{\sigma\sqrt{2\pi}} e^{\left(-0.5\left(\frac{t-\tau_i}{\sigma}\right)^2\right)} \tag{3.33}$$

The width and thus the value of σ of the Gaussian signal should be small to stay close to the Dirac impulse. However, the narrower the impulse is, the larger is its maximal value. For a reasonable $\sigma = 1 \cdot 10^{-12}s$ the maximum voltage would be $4 \cdot 10^{11}V$, which causes problems in the simulator. Therefore, the scaling coefficient θ is introduced. After the simulation this scaling is undone (see Equation (3.35)).

The output voltage $V_{vcvs}(t)$ of the voltage controlled voltage source is the sum of the voltage differences over all resistors:

$$V_{vcvs}(t) = \sum_i V_{R'_i}(t) - V_{R_i}(t) \tag{3.34}$$

All resistors have the same value which results in no change of the input signal of the stage as all the voltage differences are zero. At the output the sensitivities of the voltage to the resistor values can be computed. For each time τ_i there are two resistors R'_i and R_i where R'_i is fixed and R_i is the resistor the sensitivity analysis is performed to. The sensitivity analysis performed by the simulator yields the sensitivity $\frac{\partial V(t)}{\partial R_i}\Big|_{t=t_i}$. Let R_i be the resistor connected to the approximated Dirac impulse at time τ_i. Using the equations of a voltage divider, $h(t_j, \tau_i)$ can be obtained by:

$$h(t_j, \tau_i) \approx 2 \left.\frac{\partial V(t)}{\partial R_i}\right|_{t=t_j} \frac{R'_i}{\theta} \tag{3.35}$$

With this sampled transfer function the integration in 3.32 can be computed using means of numerical integration like Simpson rule [Jen69].

The weakness of the proposed approach are numerical problems. The approximation of the Dirac impulse by a Gaussian function causes these problems. The σ should be low to keep the signal narrow and to obtain a better match with the ideal Dirac impulse. However, this signal has to be processed by the analog simulator and for faster changing signals, the accuracy settings have to be chosen higher. This leads to higher computation time. Even with extremely small timesteps and rigid solver settings the results are not satisfying while the runtime becomes unacceptably large. Therefore, a better possibility is described in the following, which does not rely on extremely accurate approximations of Dirac impulses.

Direct Computation of the Variations at the Output

In the previous section, the voltage sources of the voltage dividers were approximations of Dirac impulses. This caused numerical problems. In the following, the same auxiliary circuit is used but instead of approximated Dirac impulses, other signals are used to avoid the numerical problems.

Equation (3.21) shows the canonical sum using time dependent sensitivities $\chi_\nu(t)$. During SSTA, these sensitivities are in a sampled form but nevertheless they can be used as sampled continuous functions. Each $\chi_\nu(t)$ represents the influence of parameter p_ν on the voltage at time t at the input. These functions are now

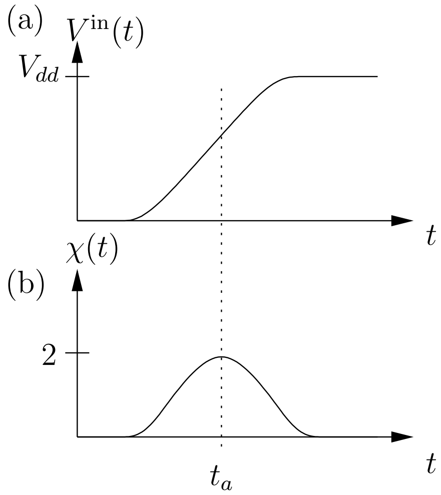

Figure 3.7: Input signal $V^{\text{in}}(t)$(a) and time dependent sensitivity $\chi(t)$(b)

used as input voltages of the voltage sources:

$$V_\nu(t) = \frac{\chi_\nu(t)}{\theta} \tag{3.36}$$

The normalization factor θ is necessary to obtain a reasonable voltage value that does not cause problems in the simulator because of too large or too small values.

This procedure is illustrated using an example. Consider the input waveform $V^{\text{in}}(t)$ in Figure 3.7(a) and the sensitivity $\chi(t)$ of the input voltage to one transistor parameter, e.g. *vth0* (b). In this example, the input voltage at time t_a changes by 2mV if *vth0* changes by 1mV as the sensitivity is 2. In the nominal case, the nominal waveform is not altered, as the variation of the parameters is zero. The question is: By how much does the output voltage at time t_i change if the parameters are altered? The sensitivity signal $\chi(t)$ is added with a factor k to the input waveform and the resulting change in the output voltage has to be determined. Using finite differences, the factor k can be chosen small and the resulting output voltage at time t_i will change by $\Delta V^{\text{out}}(t_i)$. The sensitivity is obtained by the quotient $\Delta V^{\text{out}}(t_i)/k$. The problem using finite differences is that the number of parameters and thus the number of signals $\chi_\nu(t)$ can be large resulting in a high number of simulations. As before, the computation can be transferred to the analog simulator using the voltage dividers. The resulting circuit for one parameter is shown in Figure 3.8. Altering the resistors R by ΔR would change the voltage at the input of the voltage adder and results in the superposition of $\chi(t)$ and $V^{\text{in}}(t)$. Thus, it is the same effect as using a small factor k for $\chi(t)$ directly as shown before. The difference is, that now it is possible to let the sensitivities to the resistor R be computed by the simulator.

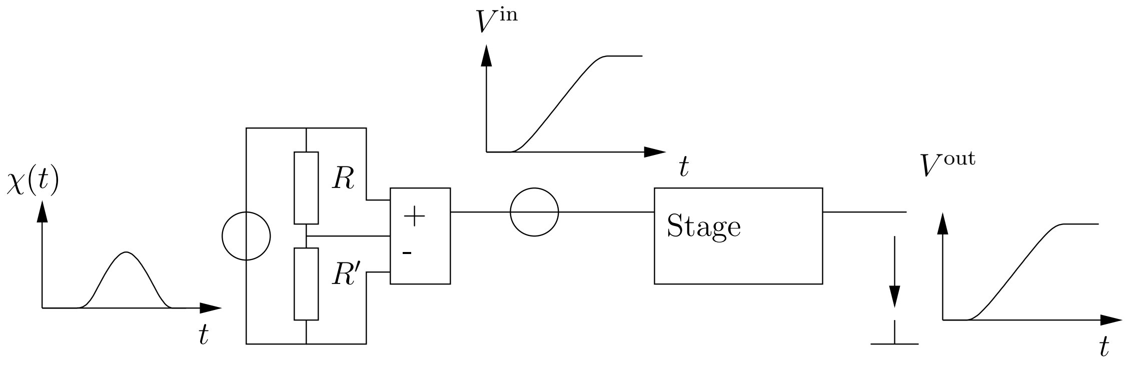

Figure 3.8: Auxiliary circuit for a single input parameter.

As before, these sensitivities of the output voltage to the resistors $\partial V^{\text{out}}/\partial R_\nu$ are obtained. The sensitivities of the output voltage to the parameters at the input can then be computed as

$$\chi_\nu^{\text{out}}(t) = 2\frac{R'_\nu}{\theta}\frac{\partial V^{\text{out}}(t)}{\partial R_\nu} \tag{3.37}$$

For more than one parameter the auxiliary circuit becomes larger. One voltage divider and one voltage source is needed for each parameter influencing the input waveform. Using the described procedure it is possible to efficiently compute the needed sensitivities $\chi_\nu^{\text{out}}(t)$ to set up the canonical sum for the voltages at the output of the stage.

Addition of Input Variations and Stage Parameter Variations

In this section the last of the three steps of the analysis of one stage will be described. After the calculation of the influence of the parameters internal to the stage and the influence of the variation at the input of the stage, these two contributions have to be added. It is crucial to note that the canonical sums of the input variation and stage parameter variation can have common variables. This results in correlation between the two contributions and this correlation has to be considered. For that purpose the notation has to distinguish between input parameters p_i^{in} and stage parameters p_i^{stage}.

Let $\mathcal{P}^{\text{in}}$ and $\mathcal{P}^{\text{stage}}$ be the set of parameters at the input and internal to the stage respectively. In case of only local variations, the two sets are disjoint ($\mathcal{P}^{\text{in}} \cap \mathcal{P}^{\text{stage}} = \emptyset$) as local parameters influencing previous cells will never influence the current cell. Considering only global variations leads to the equality of the two sets ($\mathcal{P}^{\text{in}} = \mathcal{P}^{\text{stage}}$). The value of the sensitivity at the output $\chi_\nu^{\text{out}}(t)$ can then be computed by

$$\chi_\nu^{\text{out}}(t) = \begin{cases} \chi_\nu^{\text{tr}}(t) & : \; p_\nu \in \mathcal{P}^{\text{in}} \wedge p_\nu \notin \mathcal{P}^{\text{stage}} \\ \chi_\nu^{\text{stage}}(t) & : \; p_\nu \notin \mathcal{P}^{\text{in}} \wedge p_\nu \in \mathcal{P}^{\text{stage}} \\ \chi_\nu^{\text{tr}}(t) + \chi_\nu^{\text{stage}}(t) & : \; p_\nu \in \mathcal{P}^{\text{in}} \wedge p_\nu \in \mathcal{P}^{\text{stage}} \\ 0 & : \; p_\nu \notin \mathcal{P}^{\text{in}} \wedge p_\nu \notin \mathcal{P}^{\text{stage}} \end{cases} \tag{3.38}$$

where $\chi_\nu^{\text{tr}}(t)$ is the sensitivity of the output voltage to the input parameter p_ν^{in} from Equation (3.37) and $\chi_\nu^{\text{stage}}(t)$ is the sensitivity of the output voltage to the stage parameter p_ν^{stage} from Equation (3.18). The coefficients $\chi_\nu^{\text{out}}(t)$ now contain the influence of the input variation and the variations of the devices belonging to the current stage. The output waveform variation is thus given by the canonical sum:

$$\delta V^{\text{out}}(t) = \sum_{\nu=1}^{N} \chi_\nu^{\text{out}}(t) \delta p_\nu \tag{3.39}$$

3.4.3 Propagation of Variations through a Path

After the analysis of one stage, the propagation through an entire path will be described in the following. The path is traversed stage by stage and for each stage the three basic computations are executed: Influence of the parameter variation of the current stage on the output, influence of the variations of the input waveform on the output and addition of both influences.

For each stage a transistor level SPICE simulation is performed to obtain the exact output waveform for the nominal case. In addition to this, the sensitivities discussed in the previous section are derived by an adjoint network analysis, which is implemented in the employed simulator. The output waveform and the obtained variations serve as input for the next stage. This is repeated until the end of the path is reached. The following describes the iteration process in more detail.

First Stage: The non-varying input signal at the input node of the first cell is given. The corresponding first stage consists of the first cell itself plus all relevant cells and interconnect elements in the fanout that have an influence on the output waveform of the first cell under consideration. The nominal simulation yields the nominal output waveform of the first stage, $V_{\text{nom}}^{\text{out}}(t) \equiv V_{\text{nom}}^{(1)}(t)$, for given points in time. Additionally, it explicitly derives the time derivatives $\partial V_{\text{nom}}^{(1)}(t)/\partial t$ of the output signal at these points in time. As the input signal is fixed, the output waveform of the first stage may only vary due to variations of the stage-internal transistor parameter variations $\delta p_\nu^{\text{stage}}$. In addition to the nominal output waveform $V_{\text{nom}}^{(1)}(t)$, an adjoint network analysis within the same simulation run derives the output voltage sensitivities $\partial V_{\text{nom}}^{(1)}(t)/\partial\delta p_\nu^{\text{stage}}$ and the delay sensitivity $\partial\tau/\partial p_\nu^{\text{stage}}$ with respect to these transistor parameter variations. Thus the output waveform deformation according to Equation (3.10) as well as the time shift variation $\delta\tau^{(1)}$ given by Equation (3.15) are explicitly known.

Stage i ($i = 2, 3, \dots$): The intermediate stage i consists of cell i of the path, plus all relevant cells and interconnect elements in its fanout, which have an influence on the resulting output waveform of cell i. The input signal of this stage is the output signal of the previous stage available from the previous simulation step $i-1$. The output waveform of the stage under consideration may vary due to variations of the stage-internal transistor parameter variations $\delta p_\nu^{\text{stage,i}}$ as well as the variation of the input signal $\delta V^{(i-1)}(t)$ parametrized by the linear sensitivity coefficients calculated in the previous step. Similar to the previous step, the nominal simulation for stage i produces the nominal output waveform $V_{\text{nom}}^{(i)}(t)$ and the time derivative $\partial V_{\text{nom}}^{(i)}(t)/\partial t$. In addition to this, the adjoint network analysis yields the output voltage sensitivities and the delay sensitivity with respect to the varying transistor parameters that show up in Equation (3.10) for this stage. The variation parameters of the previous stage are now the input variation variables of stage i, $\delta p_\nu^{\text{stage,i}-1} = \delta p_\nu^{\text{in,i}}$, while the internal parameters of stage i and its receivers define the new stage internal variations $\delta p_\nu^{\text{stage,i}}$. The output time shift variation of the previous stage becomes the input time shift variation of the current stage, $\delta\tau^{\text{out,i}-1} = \delta\tau^{\text{in,i}}$. As before, the simulation run of step i yields the nominal output waveform, now of the stage i, plus all the sensitivity quantities in Equation (3.10) for this stage.

Final Stage: The approach is iterated until the last stage of the given path has been simulated. As a final result, the nominal output waveform of the path is available, plus its variation properties as a function of all the transistor parameter variations as given by Equation (3.18). The output variation is parametrized as a time shift plus an additional waveform deformation. These variations are represented by canonical sums. In most cases other quantities like the standard deviation of the delay are more relevant. Therefore, the canonical sums have to be converted as described in Section 3.5.

3.4.4 Lumping of Local Variations

The previous section described the propagation of random variables through a path using canonical sums of transistor parameters. These canonical sums consist of all transistor parameters which influenced the devices on the path analyzed so far. However, the number of variables can become very large for longer paths if local variations are considered. The reason is that each device has its own set of independent random variables and each gate is composed of several of these devices. E.g., if each gate consists of 20 transistors and three local variables are considered, this leads to a growth of the canonical sum of 60 variables per stage. Global variations do not cause this increase of variables as all devices share the same global random variables.

Even though the local variables cause significant overhead, this information is not used at the end of the path. The variables are independent and thus no other device will ever share one of the local variables. In order to eliminate this ineffectiveness, the local variables $\delta\zeta_\mu^{(i)}$ can be lumped to a single independent variable after the computation of each stage [ZCHpC06]. Thus, $\sum_{\mu=1}^{L} \alpha_{\mu,\text{local}}^{(i)} \, \delta\zeta_\mu^{(i)}$ can be replaced by the simple lumped expression $\vartheta^{(i)} \, \delta\eta^{(i)}$ with $\vartheta^{(i)} \equiv \sqrt{\sum_{\mu=1}^{L} \alpha_{\mu,\text{local}}^{(i)\;2}}$ and a new single local variable $\delta\eta^{(i)}$ with zero mean and unit variance. Hereby, the number of local variables is reduced at each stage during the propagation of the variations.

When considering spatial correlations by using a quad-tree model, a high number of variables is needed to capture the correlations. However, the variables can be lumped once the path leads into a different partition of the respective layer. Using other correlation models, the variables will be lumped once the correlation to the other variables is negligible. This procedure keeps the number of variables in the linear expression acceptable.

The disadvantage of this simplification method is that the information of the contributions of the individual local variables is lost. This becomes especially important if the time shift variation is propagated separately and is added at the end of the circuit. The voltage variations and the time shift variation are correlated because they both are influenced by the same random variables. This correlation can be considered only by keeping the information about the individual contributions of each random variable. More mathematical and quantitative details can be found in Section 3.5.1.

3.5 Post-Processing of the Results

After the last stage is processed, the results have to be interpreted. Therefore, some post-processing has to be done. This includes the composition of arrival time variation and voltage variations and also the computation of variance or other quantitative measures which can be used further in the design process.

3.5.1 Composition of Arrival Time Variation and Voltage Variations

During the path-based analysis the voltage variations were propagated separately from the variation of the arrival time. In order to obtain a meaningful result, these two contributions have to be composed again. Note that the voltage variations also contribute to the arrival time variation at the end of the path.

As mentioned earlier (see Equation (3.11)), the voltage variations can be transformed into variation of the time at a fixed voltage. This is done with the voltage variations of the waveform. The voltage variations δV_i are obtained for certain times t_i. These variations can be transformed into variations of time $\delta\tilde{t_i}$ at fixed voltages V_i, which are solely caused by the variation of the voltage by the following equation

$$\delta\tilde{t_i} = -\sum_{\nu=1}^{N} \chi_\nu(t_i) \left(\left.\frac{\partial V(t)}{\partial t}\right|_{t=t_i} \right)^{-1} \delta p_\nu \tag{3.40}$$

This contribution is added to the time shift variation $\delta\tau$ which is propagated separately through the path leading to the variations of time δt_i including both factors – voltage variations and time shift variation:

$$\delta t_i = \sum_{\nu=1}^{N} \left[\left(\vartheta_\nu - \chi_\nu(t_i) \left(\left.\frac{\partial V(t)}{\partial t}\right|_{t=t_i} \right)^{-1} \right) \delta p_\nu \right] \tag{3.41}$$

After these computations the canonical form for the variation of the time values is available as

$$\delta t_i = \sum_{\nu=1}^{N} \xi_{i,\nu}\, \delta p_\nu \tag{3.42}$$

This representation is the final result of the statistical timing analysis. It is a sum of independent random variables. The sensitivities $\xi_{i,\nu}$ can be used in the design process as they reveal the influence of the various process parameters on the timing.

3.5.2 Computation of the Variance with Lumped Parameters

In order to evaluate the accuracy of the proposed method, the result should be mapped to a quantitative measure, which can be compared to a reference like Monte

Carlo analysis. This quantitative measure can be the variance $\sigma^2_{\delta t_i}$ of the time variation. The mean value is zero, because only the variation around the nominal value is considered. As the random variables are independent, the variance can be computed by

$$\sigma^2_{\delta t_i} = \sum_{\nu=1}^{N} \xi^2_\nu \, \sigma^2_{p_\nu} \tag{3.43}$$

Using the lumping technique described in Section 3.4.4 causes some problems as the time shift variation and the voltage variations are correlated. These correlations can not be considered by the canonical sum as the parameters are lumped after each stage. In Equation (3.41) it is assumed that ϑ_ν and χ_ν are the coefficients to the same independent parameter p_ν. However, if the local variables are lumped, the random variable p_{local} contains all local variations. Such a random variable exists in the canonical sums of the voltage variations as well as in the canonical sum of the time shift variation and these variables are correlated. When the variation in time caused by the voltage variation is added to the time shift variation propagated separately, this correlation causes an error in the result.

The relative error gets smaller for longer paths because of two reasons: Firstly, the relative influence of local variations becomes smaller by $1/\sqrt{n}$ with n being the number of stages in the path. The second reason is that the coefficients for the local parameters for the time shift variation get larger for longer paths as each stage contributes while the coefficients of the voltage variation stays small. This effect reduces the value of the covariance between the two parameters and thus the error caused by neglecting that variance. This should be clarified by the following. Let $A = \mathbf{a}^T\mathbf{p}$ be the canonical sum for the time shift variation in vector notation and $B = \mathbf{b}^T\mathbf{p}$ the variation of the time caused by the voltage variation, again in vector notation. Note that the local parameters $\mathbf{p}$ are the same for A and B and thus, A and B are correlated. As in the chapters before, the random variables are zero mean. Consider the variance of $A + B$

$$\operatorname{var}(A + B) = \operatorname{var}(A) + \operatorname{var}(B) + 2\operatorname{cov}(A, B) \tag{3.44}$$

While the variance of A is given as

$$\operatorname{var}(A) = \sum_{\nu=1}^{N} a^2_\nu \, \sigma^2_{p_\nu} \tag{3.45}$$

the covariance $\mathrm{cov}(A,B)$ can be computed by definition by

$$\begin{aligned}\mathrm{cov}(A,B) &= E[A\,B] = E[\mathbf{a}^T\mathbf{p}\,\mathbf{b}^T\mathbf{p}] = E[\mathbf{a}^T\mathbf{p}\,\mathbf{p}^T\mathbf{b}] = \\ &= \mathbf{a}^T E[\mathbf{p}\mathbf{p}^T]\mathbf{b} = \mathbf{a}^T\mathbf{C_p}\mathbf{b} \qquad (3.46)\end{aligned}$$

where $\mathbf{C_p}$ is the covariance matrix of $\mathbf{p}$. Because all random variables in $\mathbf{p}$ are independent of each other, the covariance matrix $\mathbf{C_p}$ simplifies to the diagonal matrix $\mathbf{C_p} = \mathrm{diag}(\sigma_{p_1}^2, \sigma_{p_2}^2, \dots, \sigma_{p_N}^2)$ and Equation (3.46) becomes

$$\mathrm{cov}(A,B) = \sum_{\nu=1}^{N} a_\nu\,\sigma_{p_\nu}^2\,b_\nu \qquad (3.47)$$

When the lumped variables are added without considering the correlation, the term $2\,\mathrm{cov}(A,B)$ in Equation (3.44) is neglected. The relative error caused by this can be computed as

$$\begin{aligned}\mathrm{err} &= \frac{2\,\mathrm{cov}(A,B)}{\mathrm{var}(A)+\mathrm{var}(B)+2\,\mathrm{cov}(A,B)} \qquad (3.48)\\ &= \frac{2\,\sum_{\nu=1}^{N} a_\nu\,\sigma_{p_\nu}^2 b_\nu}{\sum_{\nu=1}^{N} a_\nu^2\,\sigma_{p_\nu}^2 + \sum_{\nu=1}^{N} b_\nu^2\,\sigma_{p_\nu}^2 + 2\,\sum_{\nu=1}^{N} a_\nu\,\sigma_{p_\nu}^2 b_\nu}\end{aligned}$$

This error function has a maximum at $\mathbf{a} = \mathbf{b}$ and decreases for larger differences between the elements of $\mathbf{a}$ and $\mathbf{b}$. As mentioned earlier the coefficients of the time shift variation will get larger during the propagation while the variation of the time caused by the voltage variations stay small because the time shift is subtracted at each stage. Therefore, the lumping of local parameters causes less error for longer paths. For short paths, all local variables can be kept individually as the number of local variables is small for a small number of devices in the path.

Table 3.1 shows the error caused by lumping local variations for chains of 4-input NAND gates. The library gate with the maximal driver strength was chosen as this gate comprises a large number of transistors and thus, a large number of random variables for local variations is needed. For each gate 128 random variables are needed to represent the local variations. The table shows how the error is smaller for longer paths and how the speedup achieved by lumping rises for longer paths. This significant increase is caused by the increasing number of random variables (N in Equation (3.22)) and thus, the increasing number of voltage dividers in the auxiliary circuit in Figure 3.5 on page 49 and sensitivity computations as in Equation (3.37).

path length	**error**	**speedup**	**#RVs no lumping**	**#RVs with lumping**
3	0.4%	3x	392	9
6	0.2%	7x	776	9
10	0.03%	14x	1288	9

Table 3.1: Implications of lumping local variables for paths of 4-input NAND gates.

The computations above aim for the determination of the variance of the time at particular voltage crossing points. If Gaussian distributions are assumed, the entire distribution is determined by the mean – which is zero in this case – and the variance. For other distributions this is not always the case so the entire distribution can be obtained by sampling the p_ν and evaluating (3.43) for each sample. This yields the exact distribution and other quantities than just mean and sigma can be derived. The following chapter provides some data of the results of the proposed method of various circuit examples.

Chapter 4

Results

4.1 Implementation

The proposed method was implemented as a set of Python scripts and executed on a Linux machine. After the evaluation of the method, the program was ported to integrate into an industrial framework using the in-house tool PDC. It was executed on a single core of an Intel(R) Xeon(R) E5345 2.33GHz.

4.2 Setup of Monte Carlo Analysis

Equation (3.41) describes how the crossing time is influenced by a change of transistor parameters. In other words it is the linearization of the mapping function of transistor parameters to crossing times. As long as the changes of transistor parameters remain small enough, this linearization can be used to compute the change of a crossing time resulting from a change in a transistor parameter. This information can be passed back to the designer in order to optimize the design, e.g. for yield maximization. This sensitivity information can also be used to map the statistical variations of the transistor parameters to variations of the crossing times as described earlier. The focus of this work is not the design optimization but the statistical analysis of the design. Therefore, the accuracy of the statistical results are evaluated in the following and the results derived by the proposed method are compared to Monte Carlo analysis.

The Monte Carlo analysis works by sampling the parameter space according to the given distributions. In this case the distributions of the transistor parameters are Gaussian. The statistical process data is provided by a manufacturer of integrated circuits. The considered parameters were the global parameters vth0 (threshold voltage), xl (gate length), toxe (electrical gate equivalent oxide thickness), and toxp (physical gate equivalent oxide thickness). The local parameters were delvto (local threshold variation) and factu0 (mobility multiplication factor). These parameters were selected as they showed the most significance in a wide range of simulations. The number of samples has to be sufficiently large and has been set to 10,000 to guarantee a reliable accuracy. For each of these samples of the transistor parameter space, the path is simulated by a SPICE simulator and the result is saved. This result can be the output waveform or numerical measures like the crossing times of the same voltage levels as used during the SSTA. Before evaluating the statistical results, the accuracy of the waveforms should be examined.

4.3 Influence and Accuracy of Waveforms

As mentioned earlier, the time shift variation is propagated separately. For each stage the linear sensitivity of the stage delay to the stage parameters is obtained and the result is added to the time shift variation at the input. The result should be compared to existing methods. However, the models in these existing methods show many different sources of inaccuracies. The modeling inaccuracies are hard to evaluate as most models show no error for at least one operating condition. In the following, the influence of waveform variation should be examined and thus, all other sources of inaccuracies should be excluded.

Even though only arrival time and slope are considered in classical STA, the library cells are characterized using standard waveforms as inputs. If the waveforms occurring in the actual test case are equal to these standard waveforms, no error is introduced by considering only arrival time and slope. The same applies for the receiver modeling. In order to exclude inaccurate interconnect and gate input modeling, the best result possible should be assumed. Another possible source of inaccuracies is the interpolation of previously characterized look-up tables. To exclude this influence as well, it has to be assumed that the actual values for input slope and output load are exactly one of the values of the cell characterization.

For state-of-the-art SSTA, the canonical sum of the delay of the current gate has to be obtained from characterization data. If the characterization has been done for the same conditions as in the actual circuit, the results are equal to the canonical sum obtained by adjoint networks sensitivity analysis. By implementing existing methods for STA and SSTA, arbitrary errors can be introduced by unreasonable choice of model parameters. Thus, the result of such an implementation would not be meaningful for a fair comparison.

The only way to show the influence of waveform variation or advantage of considering this variation is to assume highest possible accuracy of existing methods. This is achieved here by using stage-based analog simulations for the existing methods as well. The only difference is that for the existing methods, only the delay variation is propagated while for the method proposed in this work, the variation of the entire waveform is propagated. For each instance of a cell the delay and the canonical sum of the delay variation is computed by analog simulation. This is the best-case accuracy of classical, table-based methods. Obviously, the analog simulations slow down the classical methods by orders of magnitude but it is necessary to quantify the influence of waveform variation.

In standard SSTA methods, analog simulations are only used during library characterization for various different values of input slope and output load and one standard waveform. Here, the characterization step is performed for each instance of a cell in the circuit individually considering the actual input waveform shape including the slope and the output load. This ensures the perfect match of the model to this particular cell instance and the only error that is introduced derives from not considering the waveform variation.

Figure 4.1 shows output waveforms of a chain of 10 minimal inverters with simple RC interconnects. The values for load and slope lie within the library characterization data of the cells. One arbitrary run of the Monte Carlo analysis was selected to show the difference of the resulting waveforms. The error caused by neglecting the waveform variation and only propagating the arrival time variation is shown. It can be seen that neglecting the waveform variation causes a significant deviation from the result of the SPICE simulation.

Some more output waveforms are shown in Figure 4.2. To illustrate the accuracy of the proposed method, seven samples of the Monte Carlo simulation were arbitrarily selected from all 10,000 Monte Carlo samples. For each sample, the waveform which results from full-path SPICE simulation is compared to the waveform which is obtained by the proposed method. The figure shows that the results of the proposed method are close to the results of full-path SPICE simulation.

4.4 Statistical Results

It was shown that the sensitivity coefficients in the canonical sum can be used to compute the output waveform from a given change of parameters. Now, the statisti-

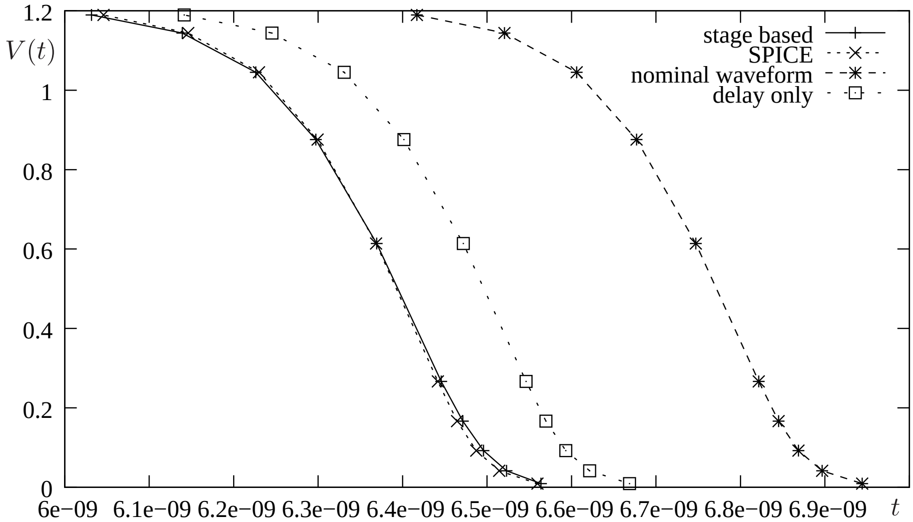

Figure 4.1: Output waveform for the nominal case and one Monte Carlo run obtained by SPICE simulation, the proposed method, and the proposed method only considering the delay variation.

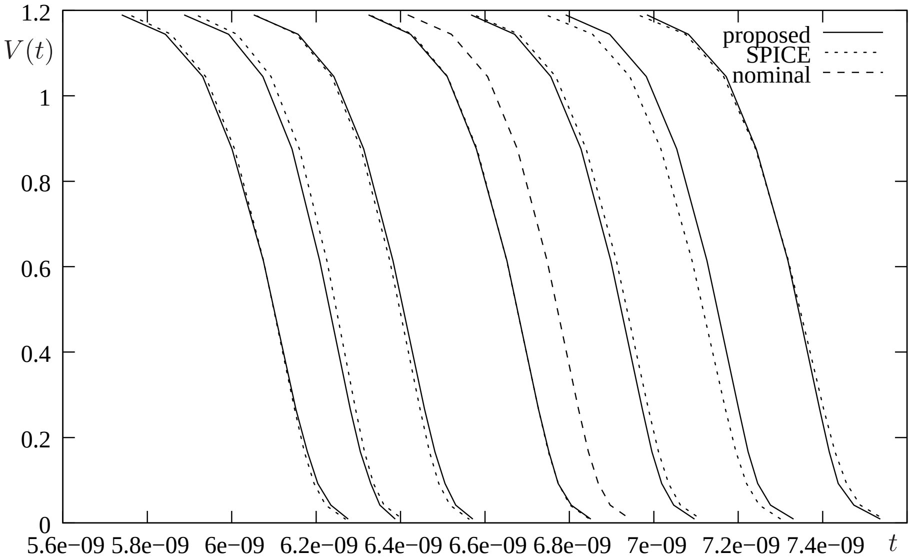

Figure 4.2: Output waveforms from the proposed method and from SPICE simulation for arbitrary samples of the Monte Carlo simulation

cal quantities should be evaluated against the result of Monte Carlo analysis. After the simulation of all samples, the statistical properties like mean and variance of the crossing times can be computed. For any of the crossing time variations the result of the proposed SSTA can be compared to the result of the Monte Carlo analysis. For the sake of a simplified presentation, only one of the crossing time variations is selected. The most significant and mostly used crossing point is the one which crosses $V_{dd}/2$ or half the voltage swing. This crossing point usually defines the arrival time at the end of the path or the delay of the path if the arrival time at the primary input is assumed to be zero. The variation of the delay is considered in the following as a simple criterion for the accuracy of the proposed method.

The proposed method was evaluated on various generic cell chains and specific paths from an industrial digital design with fully extracted post-layout parasitics. The variation modeling was based on 4 global transistor parameters and 2 local parameters per individual transistor as described in Section 4.2. The underlying technology modeling, including the statistical information, is taken from a productive 90nm technology framework. The local parameters of each transistor are independent of each other but correlations can be considered as described in Section 2.3.4. The quantity chosen for comparison is the arrival time at the output of the path. The mean of the arrival time is determined by the deterministic nominal simulation and thus not of major interest. Therefore, the accuracy of the standard deviation of the arrival time (σ_d) will serve as main quality criterion. To make sure that the Monte Carlo reference simulation itself is sufficiently accurate we used 10,000 Monte Carlo runs per testcase.

The first set of simple test cases consists of generic cell chains with 10 instances of the identical library cell in each path connected by a simple interconnect structure of 6 resistors and five capacitors between each cell. Minimally sized cells where chosen to maximize the influence of parameter variations. Table 4.1 shows the number and type of gates used in these simple test circuits, the required computation time, the computation time for a corresponding Monte Carlo simulation, and the error in the σ_d value, calculated as the relative deviation of the results from our approach compared to the Monte Carlo reference.

The next set of test cases is a collection of specific paths from an industrial digital design with extracted post-layout parasitic nets. The netlists of these paths incorporate not only the gates in the path but also the relevant receiver gates in the fanout of the respective gates. Table 4.2 shows the number of cells, the overall number of resistors (#res) and capacitors (#cap) involved, the computation time

Type	#gates	err σ_d	rt	rt(MC)
INV	10	3%	67s	$6 \cdot 10^3$s
CLK_BUF	10	3%	141s	$14 \cdot 10^3$s
MUX4	10	2%	799s	$131 \cdot 10^3$s
NAND4	10	1%	427s	$34 \cdot 10^3$s
NOR3	10	3%	374s	$32 \cdot 10^3$s

Table 4.1: Results for simple chains

#gates	#res/#cap	err σ_d	rt	rt(MC)
35	237/215	3%	923s	$614 \cdot 10^3$s
35	241/219	5%	927s	$608 \cdot 10^3$s
34	225/204	4%	907s	$536 \cdot 10^3$s
50	108/44	5%	1400s	$1.2 \cdot 10^6$s

Table 4.2: Results for paths from an industrial design

for the proposed method, the computation time needed for a full-path Monte Carlo SPICE simulation, and the error of the standard deviation of the delay σ_d derived by our method compared to the Monte Carlo simulation.

As the tables show, the difference between the σ_d obtained by the proposed method and the σ_d from the full-path Monte Carlo SPICE simulation is excellent (at most 5%) for both the generic cell chains and the real-life industrial examples, while analyzing the industrial examples using the proposed method is 600-800 times faster than the corresponding Monte Carlo simulation.

Next, the resulting probability density functions (pdfs) should be evaluated. Figure 4.3 shows the comparison of the histogram from the Monte Carlo simulation and the Gaussian pdf with the variance resulting from the proposed method for the arrival time t_a, i.e. the crossing time of $V_{dd}/2$, of the output signal of the path. The figure shows a close match between the histogram and the pdf. Thus, the assumption of Gaussianity at the output of the circuit is reasonably accurate.

Due to the separation of a path into stages, the complexity of the proposed method is linear in the number of gates in the path while the complexity for full-path Monte Carlo SPICE simulation is higher. Thus the proposed method offers a highly accurate statistical timing analysis considering the entire waveform between the cells in cases where Monte Carlo analysis is prohibitive due to runtime constraints. As such, it is ideally suited to be used as a reference for evaluating the accuracy of commercial SSTA tools, which use simplifying assumptions to achieve greater speeds.

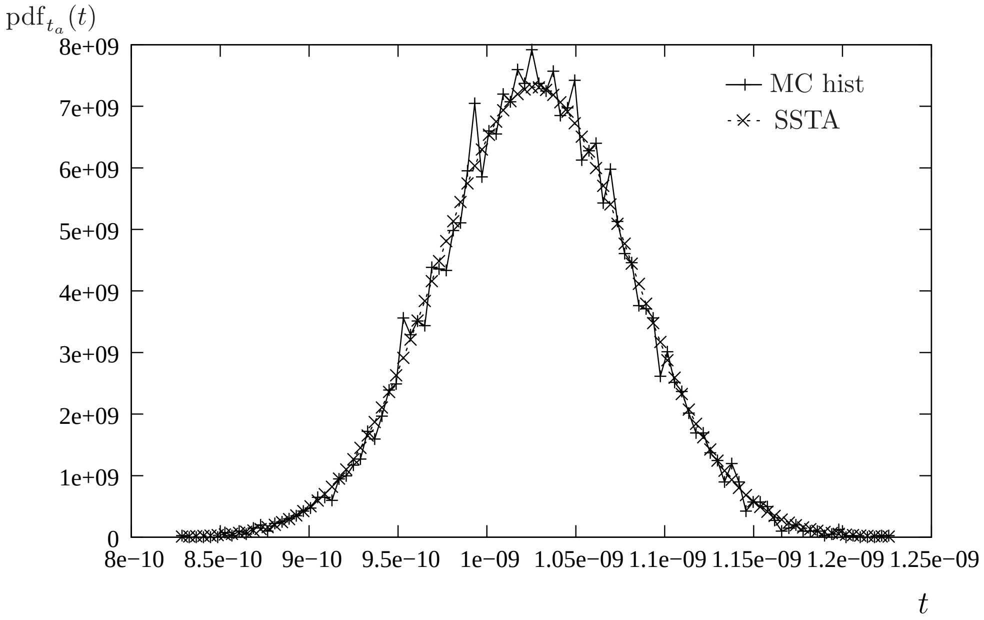

Figure 4.3: Comparison of the histogram of the Monte Carlo simulation and a Gaussian pdf with σ computed by the proposed SSTA.

The approach can also be used in a productive design environment to analyze in more detail those paths that have been identified as being the statistically most critical paths by a commercial SSTA tool.

Chapter 5

Conclusion

Timing analysis is one of the major steps in the design flow of integrated digital circuits. The downscaling of feature sizes increases the relevance of various new effects which jeopardize the accuracy of standard timing tools. Two of these effects are considered in this work:

1. The **waveform** has increasing importance on the timing behavior. Due to higher resistances of the interconnect, the waveforms differ significantly from standard shapes. Furthermore, nonlinear dynamic receivers have an influence on the waveform. It was shown that it is crucial to consider the exact shape of the waveform in order to obtain an accurate result for the timing properties of a single gate and also of an entire circuit.

2. The relevance of **parameter variations** is increasing. With shrinking feature sizes the imperfections in the manufacturing process lead to larger relative deviations from the nominal values than in earlier process generations. As these imperfections are not known during the design phase, the impact has to be modeled or estimated. Traditional corner-based methods ensure correct operation for worst-case parameter sets. These methods, however, lead to an overdesign causing increased cost and power consumption. Therefore, the device parameters have to be considered as statistical variables and the timing of a circuit has to be computed in a statistical manner.

In this work, a novel path-based method for statistical static timing analysis (SSTA) is proposed. It is based on successive analog simulations of the stages. Each stage of a path consists of the driving cell, the interconnect structure and all receiving gates connected to the output of the driving cell. The three parts of modeling each stage were proposed as:

1. The **waveform** is the entire simulator output of the previous stage. This output signal is used as an input for the current stage. The variation of the

signal is modeled by the variation of the voltage at particular points on the waveform.

2. The **driving gate** is the full transistor level netlist as defined in the library. There are no simplifying abstraction made which could cause inaccuracies due to modeling.

3. The **dynamic load** connected to the driving gate is the post-layout interconnect structure with all parasitic RC-trees. The gates connected to that interconnect structure are also contained in the receiver because using only a single capacitance for the gate input load neglects relevant effects like the Miller effect.

The procedure for analyzing a single stage and propagating the variations from the input of the stage to the output was partitioned into the following steps:

1. The **nominal waveform** is determined. The nominal waveform is determined by the analog simulation of the entire stage for nominal transistor parameter.

2. The influence of the **input waveform variation** to the output waveform is computed. This can be achieved either by estimating the linear, time variant transfer function of the cell or by applying the parameter variations at the input and computing the sensitivities on these variations. The latter is used in this work as it is numerically advantageous for the analog simulator. The computation of sensitivities to the input waveform variation is transferred to the computation of sensitivities to resistors using a auxiliary circuit connected to the input of the current stage.

3. The **sensitivities to stage parameters** are computed. The influence of parameters internal to the stage is computed by the employed analog simulator using adjoint network analysis.

4. **Addition of both influences** is performed. The variation caused by the variation of the input signal is added to the variation caused by the stage parameter variations. During this addition, the occurrence of parameters in both influences has to be considered in order to keep the correlation information correctly.

The result of the proposed method is one weighted sum of transistor parameters for each of the considered points on the output waveform of a path. With these linear expressions the random variables of the transistor parameter variations can be mapped to the variation of timing parameters such as the path delay.

The results show that the proposed method is suitable for accurate SSTA. The aim of this work is to provide a reference methodology in order to evaluate commercially available tools. Thus, it is acceptable that the runtime is too high for timing sign-off. Only a small number of paths will be analyzed and the result will be checked against the result of faster tools. Such a reference is necessary as analog Monte Carlo simulation is too time consuming for most paths. In contrast to recently published methods, this method avoids simplifications on most levels and reaches accuracy comparable to Monte Carlo simulation. Other approaches are not evaluated against analog Monte Carlo simulation of industrial designs with extracted parasitics.

The proposed method is implemented as an in-house tool at a leading semiconductor manufacturer for accurate and reliable SSTA reference.

Bibliography

[ABZ03] A. Agarwal, D. Blaauw, V. Zolotov: *Statistical timing analysis for intra-die process variations with spatial correlations*, in: *IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, pages 900–907, 2003.

[ABZV03a] A. Agarwal, D. Blaauw, V. Zolotov, S. Vrudhula: *Computation and refinement of statistical bounds on circuit delay*, in: *ACM/IEEE Design Automation Conference (DAC)*, pages 348–353, 2003.

[ABZV03b] A. Agarwal, D. Blaauw, V. Zolotov, S. Vrudhula: *Statistical timing analysis using bounds*, in: *Design, Automation and Test in Europe (DATE)*, 2003.

[ADI03] C. S. Amin, F. Dartu, Y. I. Ismail: *Weibull based analytical waveform model*, in: *IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, pages 161–168, 2003.

[ADI05] C. S. Amin, F. Dartu, Y. I. Ismail: *Weibull-based analytical waveform model*, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, volume 24(8), August 2005.

[AFP06] S. Abbaspour, H. Fatemi, M. Pedram: *Non-gaussian statistical interconnect timing analysis*, in: *Design, Automation and Test in Europe (DATE)*, 2006.

[AFP07] S. Abbaspour, H. Fatemi, M. Pedram: *Parameterized non-gaussian variational gate timing analysis*, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, volume 26(8), August 2007.

[AZB03] A. Agarwal, V. Zolotov, D. T. Blaauw: *Statistical timing analysis using bounds and selective enumeration*, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, volume 22(9), 2003.

[BBC+07] D. S. Boning, K. Balakrishnan, H. Cai, N. Drego, A. Farahanchi, K. M. Gettings, D. Lim, A. Somani, H. Taylor, D. Truque, X. Xie: *Variation*, in: *International Symposium on Quality Electronic Design*, 2007.

[BBC+08] D. S. Boning, K. Balakrishnan, H. Cai, N. Drego, A. Farahanchi, K. M. Gettings, L. Daihyun, A. Somani, H. Taylor, D. Truque, X. Xiaolin: *Variation*, IEEE Transactions on Semiconductor Manufacturing (SM), volume 21(1), pages 63–71, 2008.

[BCSS08] D. Blaauw, K. Chopra, A. Srivastava, L. Scheffer: *Statistical timing analysis: From basic principles to state of the art*, in: *IEEE Trans. on CAD of Integrated Circuits and Systems*, volume 4 of *27*, pages 589–607, 2008.

[BKN+03] S. Borkar, T. Karnik, S. Narendra, J. Tschanz, A. Keshavarzi, V. De: *Parameter variations and impact on circuits and microarchitecture*, in: *ACM/IEEE Design Automation Conference (DAC)*, June 2003.

[BM06] J. Bargfrede, M. Mirbeth: *Statische Laufzeitanalyse digitaler Schaltungen mittels Analogsimulationen*, in: *GME/ITG-Diskussionssitzung Entwicklung von Analogschaltungen mit CAE-Methoden*, pages 137–141, VDE Verlag, 2006.

[BVB05] S. Bhardwaj, S. Vrudhula, D. Blaauw: *Probability distribution of signal arrival times using bayesian networks*, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, volume 24(11), November 2005.

[BVGC06] S. Bhardwaj, S. Vrudhula, P. Ghanta, Y. Cao: *Modeling of intra-die process variations for accurate analysis and optimization of nanoscale circuits*, in: *ACM/IEEE Design Automation Conference (DAC)*, 2006.

[CC05] Y. Cao, L. T. Clark: *Mapping statistical process variations toward circuit performance variability: An analytical modeling approach*, in: *ACM/IEEE Design Automation Conference (DAC)*, 2005.

[CC07] Y. Cao, L. T. Clark: *Mapping statistical process variations toward circuit performance variability: An analytical modeling approach*, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, volume 26(10), October 2007.

[CKK+08] C. Cho, D. D. Kim, J. Kim, J.-O. Plouchart, D. Lim, S. Cho, R. Trzcinski: *Decomposition and analysis of process variability using constrained* *principal component analysis*, IEEE Transactions on Semiconductor Manufacturing (SM), volume 21(1), February 2008.

[Cla61] C. E. Clark: *The greatest of a finite set of random variables*, Operations Research, volume 9(2), pages 145–162, March 1961.

[CM06] M. Choi, L. Milor: *Impact on circuit performance of deterministic within-die variation in nanoscale semiconductor manufacturing*, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, volume 25(7), July 2006.

[CS03a] J. P. Cain, C. J. Spanos: *Electrical linewidth metrology for systematic CD variation characterization and causal analysis*, in: D. Herr (ed.), *Metrology, Inspection, and Process Control for Microlithography XVII*, volume 5038, pages 350–361, SPIE, 2003.

[CS03b] H. Chang, S. S. Sapatnekar: *Statistical timing analysis considering spatial correlations using a single pert-like traversal*, in: *IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, pages 621–625, 2003.

[CS05a] H. Chang, S. S. Sapatnekar: *Full-chip analysis of leakage power under process variations, including spatial correlations*, in: *ACM/IEEE Design Automation Conference (DAC)*, June 2005.

[CS05b] H. Chang, S. S. Sapatnekar: *Statistical timing analysis under spatial correlations*, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, volume 24(9), September 2005.

[CW03] J. Croix, M. Wong: *Blade and razor: cell and interconnect delay analysis using current-based models*, in: *ACM/IEEE Design Automation Conference (DAC)*, June 2003.

[CZNV05] H. Chang, V. Zolotov, S. Narayan, C. Visweswariah: *Parameterized block-based statistical timing analysis with non-gaussian parameters, nonlinear delay functions*, in: *ACM/IEEE Design Automation Conference (DAC)*, 2005.

[DK03] A. Devgan, C. Kashyap: *Block-based static timing analysis with uncertainty*, in: *IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, pages 607–614, 2003.

[DMQP94] F. Dartu, N. Menezes, J. Qian, L. T. Pillage: *A gate-delay model for high-speed CMOS circuits*, in: *ACM/IEEE Design Automation Conference (DAC)*, 1994.

[DR69] S. Director, R. A. Rohrer: *The generalized adjoint network and network sensitivities*, IEEE Transactions on CT, volume 16, pages 318–323, Aug 1969.

[EBSLM97] M. Eisele, J. Berthold, D. Schmitt-Landsiedel, R. Mahnkopf: *The impact of intra-die device parameter variations on path delays and on the design for yield of low voltage digital circuits*, IEEE Transactions on VLSI Systems, volume 5(4), pages 360–368, December 1997.

[FCC$^+$05] P. Friedberg, Y. Cao, J. Cain, R. Wang, J. Rabaey, C. Spanos: *Modeling within-die spatial correlation effects for process-design co-optimization*, in: *International Symposium on Quality Electronic Design*, 2005.

[FNP06] H. Fatemi, S. Nazarian, M. Pedram: *Statistical logic cell delay analysis using a current-based model*, in: *ACM/IEEE Design Automation Conference (DAC)*, 2006.

[GNDL01] A. Gattiker, S. Nassif, R. Dinakar, C. Long: *Timing yield estimation from static timing analysis*, in: *International Symposium on Quality Electronic Design*, 2001.

[GV08] A. Goel, S. Vrudhula: *Statistical waveform and current source based standard cell models for accurate timing analysis*, in: *ACM/IEEE Design Automation Conference (DAC)*, pages 227–230, June 2008.

[GvL85] G. H. Golub, C. F. van Loan: *Matrix computations*, Johns Hopkins University Press, Baltimore, 1985.

[Hit82] R. B. Hitchcock: *Timing verification and the timing analysis program*, in: *ACM/IEEE Design Automation Conference (DAC)*, pages 594–604, June 1982.

[HJ87] N. Hedenstierna, K. O. Jeppson: *CMOS circuit speed and buffer optimization*, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, volume 6(2), March 1987.

[HKO01] A. Hyvärinen, J. Karhunen, E. Oja: *Independent component analysis*, Wiley & Sons, 1st edition, 2001.

[HYO04] M. Hashimoto, Y. Yamada, H. Onodera: *Equivalent waveform propagation for static timing analysis*, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, volume 23(4), April 2004.

[ISNM08] M. Imai, T. Sato, N. Nakayama, K. Masu: *Non-parametric statistical static timing analysis: An SSTA framework for arbitrary distribution*, in: *ACM/IEEE Design Automation Conference (DAC)*, pages 698–701, 2008.

[Jen69] W. Jentsch: *Digitale Simulation kontinuierlicher Systeme*, R.Oldenbourg Verlag, München, 1969.

[JKN$^+$03] J. A. G. Jess, K. Kalafala, S. R. Naidu, R. H. J. M. Otten, C. Visweswariah: *Statistical timing for parametric yield prediction of digital integrated circuits*, in: *ACM/IEEE Design Automation Conference (DAC)*, June 2003.

[JKN$^+$06] J. A. G. Jess, K. Kalafala, S. R. Naidu, R. H. J. M. Otten, C. Visweswariah: *Statistical timing for parametric yield prediction of digital integrated circuits*, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, volume 25(11), November 2006.

[Kez06] R. C. Kezer: *Characterization Guidelines for ECSM Timing Libraries*, Silicon Integration Initiative, Inc., 9111 Jollyville Road, Austin TX 78759, 2006.

[KL01] A. Korshak, J.-C. Lee: *An effective current source cell model for VDSM delay calculation*, in: *International Symposium on Quality Electronic Design*, 2001.

[KS05] V. Khandelwal, A. Srivastava: *A general framework for accurate statistical timing analysis considering correlations*, in: *ACM/IEEE Design Automation Conference (DAC)*, 2005.

[KTV04] I. Keller, K. Tseng, N. Verghese: *A robust cell-level crosstalk delay change analysis*, in: *IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, 2004.

[Kuh07] K. Kuhn: *Reducing variation in advanced logic technologies: Approaches to process and design for manufacturability of nanoscale cmos*, IEEE International Electron Devices Meeting, December 2007.

[LCKK01] J.-J. Liou, K.-T. Cheng, S. Kundu, A. Krstic: *Fast statistical timing analysis by probabilistic event propagation*, in: *ACM/IEEE Design Automation Conference (DAC)*, pages 661 – 666, 2001.

[LHC86] K. Lakshmikumar, R. Hadaway, M. Copeland: *Characterization and modeling of mismatch in MOS transistors for precision analog design*, IEEE Journal of Solid-State Circuits SC, volume 21, pages 1057–1066, 1986.

[LLCP08] X. Li, J. Le, M. Celik, L. T. Pileggi: *Defining statistical timing sensitivity for logic circuits with large-scale process and environmental variations*, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, volume 27(6), pages 1041–1054, 2008.

[LLGP04] X. Li, J. Le, P. Gopalakrishnan, L. T. Pileggi: *Asymptotic probability extraction for non-normal distributions of circuit performance*, in: *IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, 2004.

[LTCC08] J.-H. Liu, M.-F. Tsai, L. Chen, C. C.-P. Chen: *Accurate and analytical statistical spatial correlation modeling for VLSI DFM applications*, in: *ACM/IEEE Design Automation Conference (DAC)*, pages 694–697, 2008.

[LWA06] B. N. Lee, L.-C. Wang, M. S. Abadir: *Refined statistical static timing analysis through learning spatial delay correlations*, in: *ACM/IEEE Design Automation Conference (DAC)*, 2006.

[Moo65] G. E. Moore: *Cramming more components onto integrated circuits*, Electronics, volume 38(8), April 1965.

[Moo75] G. Moore: *Progress in digital integrated electronics*, in: *International Electron Devices Meeting*, 1975.

[Nai02] S. R. Naidu: *Timing yield calculation using an impulse-train approach*, in: *Asia and South Pacific Design Automation Conference*, 2002.

[Nas00] S. R. Nassif: *Design for variability in dsm technologies*, in: *International Symposium on Quality Electronic Design*, March 2000.

[Nas01] S. R. Nassif: *Modeling and analysis of manufacturing variations*, in: *IEEE Custom Integrated Circuits Conference (CICC)*, 2001.

[NP06] S. Nazarian, M. Pedram: *Cell delay analysis based on rate-of-current change*, in: *Design, Automation and Test in Europe (DATE)*, 2006.

[OB04] M. Orshansky, A. Bandyopadhyay: *Fast statistical timing analysis handling arbitrary delay correlations*, in: *ACM/IEEE Design Automation Conference (DAC)*, pages 337–342, 2004.

[OK02] M. Orshansky, K. Keutzer: *A general probabilistic framework for worst case timing analysis*, in: *ACM/IEEE Design Automation Conference (DAC)*, pages 556 – 561, June 2002.

[OMC$^+$00] M. Orshansky, L. Milor, P. Chen, K. Keutzer, C. Hu: *Impact of systematic spatial intra-chip gate length variability on performance of high-speed digital circuits*, in: *IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, pages 62–67, 2000.

[OMC$^+$02] M. Orshansky, L. Milor, P. Chen, K. Keutzer, C. Hu: *Impact of spatial intrachip gate length variability on the performance of high-speed digital circuits*, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, volume 21(5), May 2002.

[OYO03] K. Okada, K. Yamaoka, H. Onodera: *A statistical gate-delay model considering intra-gate variability*, in: *IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, pages 908–913, 2003.

[Pap91] A. Papoulis: *Probability, Random Variables and Stochastic Processes*, McGraw-Hill, 1991.

[PRV95] L. T. Pillage, R. A. Rohrer, C. Visweswariah: *Electronic Circuit and System Simulation Methods*, McGraw-Hill, Inc., 1995.

[PSD70] P. Penfield, R. Spence, S. Duinker: *Tellegen's Theorem and Electrical Networks*, The M.I.T. Press, 1970.

[QPP94] J. Qian, S. Pullela, L. Pillage: *Modeling the "effective capacitance" for the RC interconnect of CMOS gates*, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, volume 13(12), December 1994.

[Sap04] S. Sapatnekar: *Timing*, Kluwer Academic Publishers, 2004.

[SBS+07] R. Singha, A. Balijepalli, A. Subramaniam, F. Liu, S. Nassif: *Modeling and analysis of non-rectangular gate for post-lithography circuit simulation*, in: *ACM/IEEE Design Automation Conference (DAC)*, pages 823–828, 2007.

[SFS+99] T. H. Smith, S. J. Fang, J. A. Stefani, G. B. Shinn, D. S. Boning, S. W. Butler: *On-line patterned wafer thickness control of chemical-mechanical polishing*, Journal of Vacuum Science & Technology A: Vacuum, Surfaces, and Films, volume 17(4), pages 1384–1390, July 1999.

[SKS08a] M. Schmidt, H. Kinzelbach, U. Schlichtmann: *Genauere Laufzeitanalyse digitaler Schaltungen durch Berücksichtigung statistischer Schwankungen der Signalformen*, in: edaCentrum (ed.), *edaWorkshop*, May 2008.

[SKS08b] M. Schmidt, H. Kinzelbach, U. Schlichtmann: *More accurate statistical timing analysis by considering waveform variations*, Technical report, Technische Universität München, Lehrstuhl für Entwurfsautomatisierung, January 2008.

[SKS08c] M. Schmidt, H. Kinzelbach, U. Schlichtmann: *Variational waveform propagation for accurate statistical timing analysis*, in: *ACM/IEEE International Workshop on Timing Issues in the Specification and Synthesis of Digital Systems*, February 2008.

[SLS+07] M. Schmidt, B. Li, W. Schneider, H. Kinzelbach, U. Schlichtmann: *Statistical timing analysis using weibull waveform modeling*, in: *International Symposium on Integrated Circuits*, September 2007.

[SS06] J. Singh, S. Sapatnekar: *Statistical timing analysis with correlated non-gaussian parameters using independent component analysis*, in: *ACM/IEEE Design Automation Conference (DAC)*, July 2006.

[SS08] J. Singh, S. S. Sapatnekar: *A scalable statistical static timing analyzer incorporating correlated non-gaussian and gaussian parameter variations*, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, volume 27(1), January 2008.

[SSA+05] A. Srivastava, S. Shah, K. Agarwal, D. Sylvester, D. Blaauw, S. Director: *Accurate and efficient gate-level parametric yield estimation considering correlated variations in leakage power and performance*, in: *ACM/IEEE Design Automation Conference (DAC)*, June 2005.

[SSB05] A. Srivastava, D. Sylvester, D. Blaauw: *Statistical Analysis and Optimization for VLSI: Timing and Power*, Springer Science+Business Media, 2005.

[TIT07] *The international technology roadmap for semiconductors*, http://www.itrs.net, 2007.

[Tur36] A. M. Turing: *On computable numbers, with an application to the Entscheidungsproblem*, in: *Proceedings of the London Mathematical Society*, 1936.

[Unb72] R. Unbehauen: *Systemtheorie*, R.Oldenbourg Verlag, München, 1972.

[VRK$^+$04] C. Visweswariah, K. Ravindran, K. Kalafala, S. Walker, S. Narayan: *First-order incremental block-based statistical timing analysis*, in: *ACM/IEEE Design Automation Conference (DAC)*, pages 331–336, 2004.

[VRK$^+$06] C. Visweswariah, K. Ravindran, K. Kalafala, S. G. Walker, S. Narayan, D. K. Beece, J. Piaget, N. Venkateswaran, J. G. Hemmet: *First-order incremental block-based statistical timing analysis*, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, volume 25(10), October 2006.

[WE93] N. H. E. Weste, K. Eshraghian: *Principles of CMOS VLSI Design*, Addison-Wesley, 1993.

[WO06] W.-S. Wang, M. Orshansky: *Path-based statistical timing analysis handling arbitrary delay correlations: Theory and implementation*, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, volume 25(12), December 2006.

[XZH07a] J. Xiong, V. Zolotov, L. He: *Robust extraction of spatial correlation*, in: *ACM/SIGDA International Symposium on Physical Design (ISPD)*, volume 26 of *4*, pages 619 – 631, April 2007.

[XZH07b] J. Xiong, V. Zolotov, L. He: *Robust extraction of spatial correlation*, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, volume 26(4), April 2007.

[YLNC08] Y. Ye, F. Liu, S. Nassif, Y. Cao: *Statistical modeling and simulation of threshold variation under dopant fluctuations and line-edge roughness*, in: *ACM/IEEE Design Automation Conference (DAC)*, pages 900–905, 2008.

[ZCH$^+$05] L. Zhang, W. Chen, Y. Hu, J. A. Gubner, C. C.-P. Chen: *Correlation-preserved non-gaussian statistical timing analysis with quadratic timing model*, in: *ACM/IEEE Design Automation Conference (DAC)*, 2005.

[ZCH$^+$06] L. Zhang, W. Chen, Y. Hu, J. A. Gubner, C. C.-P. Chen: *Correlation-preserved statistical timing with a quadratic form of gaussian variables*, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, volume 25(11), November 2006.

[ZCHpC06] L. Zhang, W. Chen, Y. Hu, C. C. ping Chen: *Statistical static timing analysis with conditional linear max/min approximation and extended canonical timing model*, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, volume 25(6), June 2006.

[ZSL$^+$05] Y. Zhan, A. J. Strojwas, X. Li, L. T. Pileggi, D. Newmark, M. Sharma: *Correlation-aware statistical timing analysis with non-gaussian delay distributions*, in: *ACM/IEEE Design Automation Conference (DAC)*, 2005.

[ZXA$^+$07] V. Zolotov, J. Xiong, S. Abbaspour, D. J. Hathaway, C. Visweswariah: *Compact modeling of variational waveforms*, in: *IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, 2007.

List of Figures