Chapter 2
Random Process Variation in Deep-Submicron CMOS

One of the most notable features of nanometer scale CMOS technology is the increasing magnitude of variability of the key parameters affecting performance of integrated circuits [1]. Although scaling made controlling extrinsic variability more complex, nonetheless, the most profound reason for the future increase in parameter variability is that the technology is approaching the regime of fundamental randomness in the behavior of silicon structures where device operation must be described as a stochastic process. Electric noise due to the trapping and de-trapping of electrons in lattice defects may result in large current fluctuations, and those may be different for each device within a circuit. At this scale, a single dopant atom may change device characteristics, leading to large variations from device to device [2]. As the device gate length approaches the correlation length of the oxide-silicon interface, the intrinsic threshold voltage fluctuations induced by local oxide thickness variation will become significant [3]. Finally, line-edge roughness, i.e., the random variation in the gate length along the width of the channel, will also contribute to the overall variability of gate length [4]. Since placement of dopant atoms introduced into silicon crystal is random, the final number and location of atoms in the channel of each transistor is a random variable. As the threshold voltage of the transistor is determined by the number and placement of dopant atoms, it will exhibit a considerable variation [3]. This leads to variation in the transistors' circuit-level properties, such as delay and power [5]. Predicting the timing uncertainty is traditionally done through corner-based analysis, which performs static timing analysis (STA) at multiple corners to obtain the extreme-case results. In each corner, process parameters are set at extreme points in the multidimensional space. As a consequence, the worst-case delay from the corner-based timing analysis is over pessimistic since it is unlikely for all process parameters to have extreme values at the same time. Additionally, the number of process corners grows exponentially as the number of process variations increases.

Recently, statistical STA (SSTA) has been proposed as a potential alternative to consider process variations for timing verification. In contrast to static timing analysis, SSTA represents gate delays and interconnect delays as probability distributions, and provides the distribution (or statistical moments) of each timing

A. Zjajo, *Stochastic Process Variation in Deep-Submicron CMOS*, Springer Series in Advanced Microelectronics 48, DOI: 10.1007/978-94-007-7781-1_2,

value rather than a deterministic quantity. When modeling process-induced delay variations, the sample space is the set of all manufactured dies. In this case, the device parameters will have different values across this sample space, hence the critical path and its delay will change from one die to the next. Therefore, the delay of the circuit is also a random variation, and the first task of statistical timing analysis is to compute the characteristics of this random variation. This is performed by computing its probability-distribution function or cumulative-distribution function (CDF).

Alternatively, only specific statistical characteristics of the distribution, such as its mean and standard deviation, can be computed. Note that the cumulative-distribution function and the probability-distribution function can be derived from one another through differentiation and integration. Given the cumulative-distribution function of circuit delay of a design and the required performance constraint the anticipated yield can be determined from the cumulative-distribution function. Conversely, given the cumulative-distribution function of the circuit delay and the required yield, the maximum frequency at which the set of yielding chips can be operated at can be found.

In addition to the problem of finding the delay of the circuit, it is also key to achieve operational robustness against process variability at the expense of a higher energy consumption and larger area occupation [6]. Technology scaling, circuit topologies, and architecture trends have all aligned to specifically target low-power trade-offs through the use of fine-grained parallelism [7], near-threshold design [8], V_{DD} scaling and body biasing [9]. Similarly, a cross-layer optimization strategy is devised for variation resilience, a strategy that spans from the lowest level of process and device engineering to the upper level of system architecture. Simultaneous circuit yield and energy optimization with key parameters (supply voltage V_{DD} and supply to threshold voltage ratio V_{DD}/V_T) is a part of a system-wide strategy, where critical parameters that minimize energy (e.g. V_{DD}/V_T) provide control mechanisms (e.g. adaptive voltage scaling) to runtime system. Yield constrained energy optimization, as an active design strategy to counteract process variation in sub-threshold or near-threshold operation, necessitates the need for statistical design paradigm to overcome the limitations of deterministic optimization schemes.

In this chapter, the circuits are described as a set of stochastic differential equations and Gaussian closure approximations are introduced to obtain a closed form of moment equations and compute the variational waveform for statistical delay calculation. For high accuracy in the case of large process variations, the statistical solver divides the process variation space into several sub-spaces and performs the statistical timing analysis in each sub-space. Additionally, a yield constrained sequential energy minimization framework applied to multivariable optimization is described.

The chapter is organized as follows: Sect. 2.1 focuses on the process variations modeled as a wide-sense stationary process and Sect. 2.2 discusses a solution of a system of stochastic differential equations for such process. In Sect. 2.3, statistical delay calculation and complexity reduction techniques are described. In Sect. 2.4,

a yield constrained sequential energy minimization framework is discussed. Experimental results obtained are presented in Sect. 2.5. Finally, Sect. 2.6 provides a summary and the main conclusions.

2.1 Modeling Process Variability

The availability of large data sets of process parameters obtained through parameter extraction allows the study and modeling of the variation and correlation between process parameters, which is of crucial importance to obtain realistic values of the modeled circuit unknowns. Typical procedures determine parameters sequentially and neglect the interactions between them and, as a result, the fit of the model to measured data may be less than optimum. In addition, the parameters are obtained as they relate to a specific device and, consequently, they correspond to different device sizes. The extraction procedures are also generally specialized to a particular model, and considerable work is required to change or improve these models.

For complicated IC models, parameter extraction can be formulated as an optimization problem. The use of direct parameter extraction techniques instead of optimization allows end-of-line compact model parameter determination. The model equations are split up into functionally independent parts, and all parameters are solved using straightforward algebra without iterative procedures or least squares fitting. With the constant downscaling of supply voltage the moderate inversion region becomes more and more important, and an accurate description of this region is thus essential. The threshold-voltage-based models, such as BSIM and MOS 9, make use of approximate expressions of the drain-source channel current I_{DS} in the weak inversion region (i.e., subthreshold) and in the strong-inversion region (i.e., well above threshold). These approximate equations are tied together using a mathematical smoothing function, resulting in neither a physical nor an accurate description of I_{DS} in the moderate inversion region (i.e., around threshold). The major advantages of surface potential (defined as the electrostatic potential at the gate oxide/substrate interface with respect to the neutral bulk) over threshold voltage based models is that surface potential model does not rely on the regional approach and I–V and C–V characteristics in all operation regions are expressed/evaluated using a set of unified formulas. In the surface-potential-based model, the channel current I_{DS} is split up in a drift (I_{drift}) and a diffusion (I_{diff}) component, which are a function of the gate bias V_{GB} and the surface potential at the source (υ_{s0}) and the drain (υ_{sL}) side. In this way I_{DS} can be accurately described using one equation for all operating regions (i.e., weak, moderate and strong-inversion). The numerical progress has also removed a major concern in surface potential modeling: the solution of surface potential either in a closed form (with limited accuracy) exists or as with our use of the second-order Newton iterative method to improve the computational efficiency in MOS model 11.

The fundamental notion for the study of spatial statistics is that of stochastic (random) process defined as a collection of random variables on a set of temporal or spatial locations. Generally, a second-order stationary (wide sense stationary, WSS) process model is employed, but other more strict criteria of stationarity are possible. This model implies that the mean is constant and the covariance only depends on the separation between any two points. In a second-order stationary process only the first and second moments of the process remain invariant. The covariance and correlation functions capture how the co-dependence of random variables at different locations changes with the separation distance. These functions are unambiguously defined only for stationary processes. For example, the random process describing the behavior of the transistor length L is stationary only if there is non systematic spatial variation of the mean L. If the process is not stationary, the correlation function is not a reliable measure of codependence and correlation. Once the systematic wafer-level and field-level dependencies are removed, thereby making the process stationary, the true correlation is found to be negligibly small. From a statistical modeling perspective, systematic variations affect all transistors in a given circuit equally. Thus, systematic parametric variations can be represented by a deviation in the parameter mean of every transistor in the circuit.

We model the manufactured values of the parameters $p_i \in \{p_1,\ldots,p_m\}$ for transistor i as a random variable

$$p_i = \mu_{p,i} + \sigma_p(d_i) \cdot p(d_i, \theta) \tag{2.1}$$

where $\mu_{p,i}$ and $\sigma_p(d_i)$ are the mean value and standard deviation of the parameter p_i, respectively, $p(d_i,\theta)$ is the stochastic process corresponding to parameter p, d_i denotes the location of transistor i on the die with respect to a point origin and θ is the die on which the transistor lies. This reference point can be located, say in the lower left corner of the die, or in the center, etc. A random process can be represented as a series expansion of some uncorrelated random variables involving a complete set of deterministic functions with corresponding random coefficients. A commonly used series involves spectral expansion [10], in which the random coefficients are uncorrelated only if the random process is assumed stationary and the length of the random process is infinite or periodic. The use of the Karhunen-Loève expansion [11] has generated interest because of its bi-orthogonal property, that is, both the deterministic basis functions and the corresponding random coefficients are orthogonal [12], e.g. the orthogonal deterministic basis function and its magnitude are, respectively, the eigenfunction and eigenvalue of the covariance function. Assuming that p_i is a zero-mean Gaussian process and using the Karhunen-Loève expansion, p_i can be written in truncated form (for practical implementation) by a finite number of terms M as

$$p_i = \mu_{p,i} + \sigma_p(d_i) \cdot \sum_{n=1}^{M} \sqrt{\vartheta_{p,n}}\,\delta_{p,n}(\theta) f_{p,n}(d_i) \tag{2.2}$$

where $\{\delta_n(\theta)\}$ is a vector of zero-mean uncorrelated Gaussian random variables and $f_{p,n}(d_i)$ and $\vartheta_{p,n}$ are the eigenfunctions and the eigenvalues of the covariance matrix $\Sigma_p(d_1, d_2)$ (Fig. 2.1) of $p(d_i,\theta)$, controlled through a distance based weight term, the measurement correction factor, correlation parameter ρ and process correction factors c_x and c_y.

Without loss of generality, consider for instance two transistors with given threshold voltages. In our approach, their threshold voltages are modeled as stochastic processes over the spatial domain of a die, thus making parameters of any two transistors on the die two different correlated random variables. The value of M is governed by the accuracy of the eigen-pairs in representing the covariance function rather than the number of random variables. Unlike previous approaches, which model the covariance of process parameters due to the random effect as a piecewise linear model [13] or through modified Bessel functions of the second kind [14], here the covariance is represented as a linearly decreasing exponential function

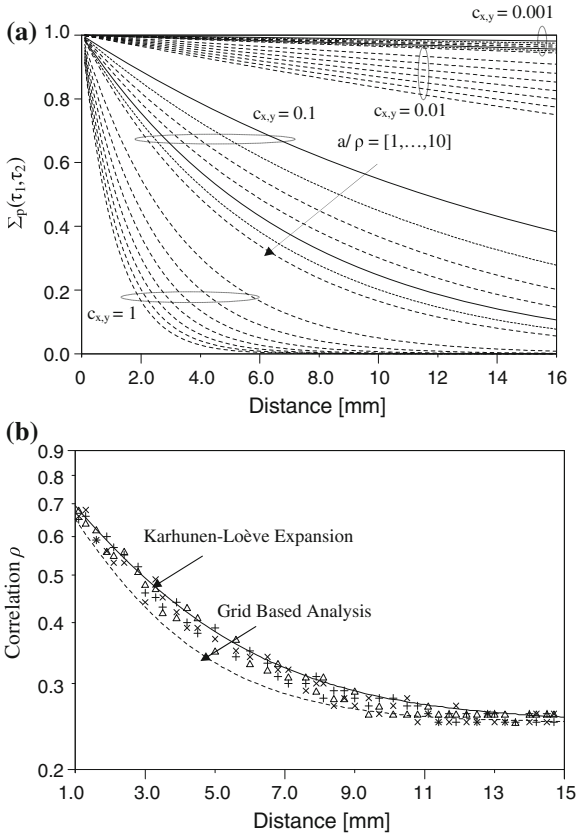

Fig. 2.1 **a** Behavior of modelled covariance functions Σ_p using $M = 5$ for $a/\rho = [1,\ldots,10]$. **b** The model fitting on the available measurement data (© IEEE 2011)

$$C_p(d_1, d_2) = \left(1 + \varsigma_{d_{x,y}}\right) \cdot \gamma \cdot \left(e^{-c_x|d_{x1}-d_{x2}|\cdot c_y|d_{y1}-d_{y2}|/\rho}\right) \tag{2.3}$$

where ς is a distance based weight term, γ is the measurement correction factor for the two transistors located at Euclidian coordinates (x_1, y_1) and (x_2, y_2), respectively, c_x and c_y are process correction factors depending upon the process maturity. For instance, in Fig. 2.1a, process correction factor $c_{x,y} = 0.001$ relates to a very mature process, while $c_{x,y} = 1$ indicates that this is a process in a ramp up phase. The correlation parameter ρ reflecting the spatial scale of clustering defined in $[-a, a]$ regulates the decaying rate of the correlation function with respect to distance (d_1, d_2) between the two transistors located at Euclidian coordinates (x_1, y_1) and (x_2, y_2).

Physically, lower a/ρ implies a highly correlated process and hence, a smaller number of random variables are needed to represent the random process and correspondingly, a smaller number of terms in the Karhunen-Loève expansion. This means that for $c_{x,y} = 0.001$ and $a/\rho = 1$ the number of, transistors that need to be sampled to assess, say a process parameter such as threshold voltage is much less than the number that would be required for $c_{x,y} = 1$ and $a/\rho = 10$ because of the high nonlinearity shown in the correlation function. To maintain a fixed difference between the theoretical value and the truncated form, M has to be increased when a increases at constant b.

In other words, for a given M, the accuracy decreases as a/b increases. Eigenvalues $\vartheta_{p,n}$ and eigenfunctions $f_{p,n}(\tau)$ are the solution of the homogeneous Fredholm integral equation of the second kind indexed on a bounded domain D. To find the numerical solution of Fredholm integral, each eigenfunction is approximated by a linear combination of a linearly decreasing exponential function. Resulting approximation error is than minimized by the Galerkin method. One example of spatial correlation dependence and model fitting on the available measurement data through Karhunen-Loève expansion is given in Fig. 2.1b. For comparison purposes, a grid-based spatial-correlation model is intuitively simple and easy to use, yet, its limitations due to the inherent accuracy-versus-efficiency necessitate a more flexible approach, especially at short to mid range distances [14]. We now introduce a model $\eta_p = f(.)$, accounting for voltage and current shifts due to random manufacturing variations in transistor dimensions and process parameters defined as

$$\eta_p = f(v, W^*, L^*, p^*) \tag{2.4}$$

where v defines a fitting parameter estimated from the extracted data, W^* and L^* represent the geometrical deformation due to manufacturing variations and p^* models electrical parameter deviations from their corresponding nominal values, e.g. altered transconductance, threshold voltage, etc. (Appendix A).

2.2 Stochastic MNA for Process Variability Analysis

Device variability effects limitations are rudimentary issues for the robust circuit design and their evaluation has been subject of numerous studies. Several models have been suggested for device variability [15–17], and correspondingly, a number of CAD tools for statistical circuit simulation [18–23]. In general, a circuit design is optimized for parametric yield so that the majority of manufactured circuits meet the performance specifications. The computational cost and complexity of yield estimation, coupled with the iterative nature of the design process, make yield maximization computationally prohibitive. As a result, circuit designs are verified using models corresponding to a set of worst-case conditions of the process parameters. Worst-case analysis refers to the process of determining the values of the process parameters in these worst-case conditions and the corresponding worst-case circuit performance values. Worst-case analysis is very efficient in terms of designer effort, and thus has become the most widely practiced technique for statistical analysis and verification. Algorithms previously proposed for worst-case tolerance analysis fall into four major categories: corner technique, interval analysis, sensitivity-based vertex analysis and Monte Carlo simulation.

The most common approach is the corners technique. In this approach, each process parameter value that leads to the worst performance is chosen independently. This method ignores the correlations among the processes parameters, and the simultaneous setting of each process parameter to its extreme value result in simulation at the tails of the joint probability density of the process parameters. Thus, the worst-case performance values obtained are extremely pessimistic. Interval analysis is computationally efficient but leads to overestimated results, i.e., the calculated response space enclose the actual response space, due to the intractable interval expansion caused by dependency among interval operands. Interval splitting techniques have been adopted to reduce the interval expansion, but at the expense of computational complexity. Traditional vertex analysis assumes that the worst case parameter sets are located at the vertices of parameter space, thus the response space can be calculated by taking the union of circuit simulation results at all possible vertices of parameter space. Given a circuit with M uncertain parameters, this will result in a 2^M simulation problem. To further reduce the simulation complexity, sensitivity information computed at the nominal parameter condition is used to find the vertices that correspond to the worst cases of circuit response. The Monte Carlo algorithm takes random combinations of values chosen from within the range of each process parameter and repeatedly performs circuit simulations. The result is an ensemble of responses from which the statistical characteristics are estimated. Unfortunately, if the number of iterations for the simulation is not very large, Monte Carlo simulation always underestimates the tolerance window. Accurately determining the bounds on the response requires a large number of simulations, so consequently, the Monte Carlo method becomes very *cpu*-time consuming if the chip becomes large. Other approaches for statistical analysis of variation-affected circuits, such as the one

based on the Hermite polynomial chaos [24] or the response surface methodology, are able to perform much faster than a Monte Carlo method at the expense of a design of an experiments preprocessing stage [25]. In this section, the circuits are described as a set of stochastic differential equations and Gaussian closure approximations are introduced to obtain a closed form of moment equations. Even if a random variable is not strictly Gaussian, a second-order probabilistic characterization yields sufficient information for most practical problems.

Modern integrated circuits are often distinguished by a very high complexity and a very high packing density. The numerical simulation of such circuits requires modeling techniques that allow an automatic generation of network equations. Furthermore, the number of independent network variables describing the network should be as small as possible. Circuit models have to meet two contradicting demands: they have to describe the physical behavior of a circuit as correct as possible while being simple enough to keep computing time reasonably small. The level of the models ranges from simple algebraic equations, over ordinary and partial differential equations to Boltzmann and Schrodinger equations depending on the effects to be described. Due to the high number of network elements (up to millions of elements) belonging to one circuit one is restricted to relatively simple models. In order to describe the physics as good as possible, so called compact models represent the first choice in network simulation. Complex elements such as transistors are modeled by small circuits containing basic network elements described by algebraic and ordinary differential equations only. The development of such replacement circuits forms its own research field and leads nowadays to transistor models with more than five hundred parameters. A well established approach to meet both demands to a certain extent is the description of the network by a graph with branches and nodes. Branch currents, branch voltages and node potentials are introduced as variables. The node potentials are defined as voltages with respect to one reference node, usually the ground node. The physical behavior of each network element is modeled by a relation between its branch currents and its branch voltages. In order to complete the network model, the topology of the elements has to be taken into account. Assuming the electrical connections between the circuit elements to be ideally conducting and the nodes to be ideal and concentrated, the topology can be described by Kirchhoff's laws (the sum of all branch currents entering a node equals zero and the sum of all branch voltages in a loop equals zero). In general, for time-domain analysis, modified nodal analysis (MNA) leads to a nonlinear ordinary differential equation or differential algebraic equation system which, in most cases, is transformed into a nonlinear algebraic system by means of linear multistep integration methods [26, 27] and, at each integration step, a Newton-like method is used to solve this nonlinear algebraic system (Appendix B). Therefore, from a numerical point of view, the equations modeling a dynamic circuit are transformed to equivalent linear equations at each iteration of the Newton method and at each time instant of the time-domain analysis. Thus, we can say that the time-domain analysis of a nonlinear dynamic circuit consists of the successive solutions of many linear

circuits approximating the original (nonlinear and dynamic) circuit at specific operating points.

Consider a linear circuit with $N + 1$ nodes and B voltage-controlled branches (two-terminal resistors, independent current sources, and voltage-controlled n-ports), the latter grouped in set B. We then introduce the source current vector $\hat{i} \in R^B$ and the branch conductance matrix $G \in R^{B\times B}$. By assuming that the branches (one for each port) are ordered element by element, the matrix is block diagonal: each 1×1 block corresponds to the conductance of a one-port and in any case is nonzero, while $n \times n$ blocks correspond to the conductance matrices of voltage-controlled n-ports. More in detail, the diagonal entries of the $n \times n$ blocks can be zero and, in this case, the nonzero off-diagonal entries, on the same row or column, correspond to voltage-controlled current sources (VCCSs). Now, consider MNA and circuits embedding, besides voltage-controlled elements, independent voltage sources, the remaining types of controlled sources and sources of process variations.

We split the set of branches B in two complementary subsets: B_V of voltage-controlled branches (v-branches) and B_C of current-controlled branches (c-branches).

Conventional nodal analysis (NA) is extended to MNA [27] as follows: currents of c-branches are added as further unknowns and the corresponding branch equations are appended to the NA system. The $N \times B$ incidence matrix A can be partitioned as $A = [A_v\ A_c]$, with $A_v \in R^{N\times Bv}$ and $A_c \in R^{N\times Bc}$. As in conventional NA, constitutive relations of v-branches are written, using the conductance submatrix $G \in R^{Bc\times Bv}$ in the form

$$i_v = Gv_v \tag{2.5}$$

while the characteristics of the c-branches, including independent voltage sources and controlled sources except VCCSs, are represented by the implicit equation

$$B_c v_c + R_c i_c + \hat{v}_c + F_c \eta = 0 \tag{2.6}$$

where B_c, R_c, $F_c \in R^{Bc\times Bc}$, $\hat{v}_c = (A^{\mathrm{T}} v_c) \in R^{Bc}$ [26] and $\eta \in R^{Bc}$ is a random vector accounting for device variations as defined in (2.4). These definitions are in agreement with those adopted in the currently used simulators and suffice for a large variety of circuits. Note that from the practical use perspective, a user may only be interested in voltage variations over a period of time or in the worst case in a period of time. This information can be obtained once the variations in any given time instance are known. By using the above notations, (2.5) and (2.6) can be written in the compact form as

$$F(q', q, t) + B(q, t) \times \eta = 0 \tag{2.7}$$

where $q = [v_c\ i_v]^{\mathrm{T}}$ is the vector of stochastic processes which represents the state variables (e.g. node voltages) of the circuit and η is a vector of wide-sense stationary processes. $B(q, t)$ is an $N \times B_c$ matrix, the entries of which are functions of the state q and possibly t. Every column of $B(q, t)$ corresponds to η, and has

normally either one or two nonzero entries. The rows correspond to either a node equation or a branch equation of an inductor or a voltage source. Equation (2.7) represents a system of nonlinear stochastic differential equations, which formulate a system of stochastic algebraic and differential equations that describe the dynamics of the nonlinear circuit that lead to the MNA equations when the random sources η are set to zero. Solving (2.7) means to determine the probability density function P of the random vector $q(t)$ at each time instant t. Formally the probability density of the random variable q is given as

$$P(q) = |\Gamma(q)|N(h^{-1}(q)|m, \Sigma) \tag{2.8}$$

where $|\Gamma(q)|$ is the determinant of the Jacobian matrix of the inverse transform $h^{-1}(q)$ with h a nonlinear function of η. However, generally it is not possible to handle this distribution directly since it is non-Gaussian for all but linear h. Therefore it may be convenient to look for an approximation which can be found after partitioning the space of the stochastic source variables η in a given number of subdomains, and then solving the equation in each subdomain by means of a piecewise-linear truncated Taylor approximation. If the subdomains are small enough to consider the equation as linear in the range of variability of η, or that the nonlinearities in the subdomains are so smooth that they might be considered as linear even for a wide range of η, it is then possible to combine the partial results and obtain the desired approximated solution to the original problem.

Let $x_0 = x(\eta_0, t)$ be the generic point around which to linearize, and with the change of variable $\xi = x - x_0 = [(q - p_0)^T, (\eta - \eta_0)^T]^T$, the first-order Taylor piecewise-linearization of (2.7) in x_0 yields

$$P(x_0)\xi' + (K(x_0) + P'(x_0))\xi = 0 \tag{2.9}$$

where $K(x) = B'(x)$, $P(x) = F'(x)$. Transient analysis requires only the solution of the deterministic version of (2.7), e.g. by means of a conventional circuit simulator, and of (2.9) with a method capable of dealing with linear stochastic differential equations with stochasticity that enters only through the initial conditions. Since (2.9) is a linear homogeneous equation in ξ, its solution, will always be proportional to $\eta - \eta_0$. We can rewrite (2.9) as

$$\xi'(x_0) = E(x_0)\xi_0 + F(x_0)\eta_0 \tag{2.10}$$

Equation (2.10) is a system of stochastic differential equations which is linear in the narrow sense (right-hand side is linear in ξ and the coefficient matrix for the vector of variation sources is independent of ξ) [28]. Since these stochastic processes have regular properties, they can be considered as a family of classical problems for the individual sample paths and be treated with the classical methods of the theory of linear stochastic differential equations. By expanding every element of $\xi(t)$ with

$$\xi_i(t) = \Gamma(t)(\eta - \eta_0) = \sum_{j=1}^{m} \alpha_{ij}(t) \cdot \eta_j \tag{2.11}$$

for m elements of a vector η. As long as $\alpha_j(t)$ is obtained, the expression for $\xi(t)$ is known, so that the covariance matrix of the solution can be written as

$$\Sigma_{\xi\xi} = \Gamma \Sigma_{\eta\eta} \Gamma^T \tag{2.12}$$

Defining $\alpha_j(t) = \left(\alpha_{1j}, \alpha_{2j}, \ldots, \alpha_{nj}\right)^T$, $F_j(t) = \left(F_{1j}, F_{2j}, \ldots, F_{nj}\right)^T$, the requirement for $\alpha(t)$ is

$$\alpha_j'(t) = E(t)\alpha_j + F(t) \tag{2.13}$$

Equation (2.13) is an ordinary differential equation, which can be solved by a fast numerical method.

2.3 Statistical Timing Analysis

Statistical static timing analysis is a potential alternative to predict the timing uncertainty due to the random process variation. In addition to the problem of finding the delay of the circuit, it is also key to improve this delay when the timing requirements are not met. Hence, deterministic STA (DSTA) methods typically report the slack at each node in the circuit, in addition to the circuit delay and critical paths. The slack of a node is the difference between the latest time a signal can arrive at that node, such that the timing constraints of the circuit are satisfied (referred to as the required time), and the actual latest arrival time of the signal at that node. Similar to the circuit delay, the slack of a node is a random variable in the SSTA formulation. Third problem associated with STA methods is latch-based sequential timing analysis, which involves multiple-phase clocks, clock-schedule verification, etc. The statistical formulation of timing analysis introduces several new modeling and algorithmic issues such as: topological correlation, spatial correlation and non-normal process parameters and nonlinear delay models.

Normal or Gaussian distributions are found to be the most commonly observed distributions for random variations, and a number of elegant analytical results exist for them in the statistics literature. However, some physical device parameters may have significantly non-normal distributions. An example of a non-normal device parameter is gate length due to the variation in depth of focus. Even if the physical device parameters are indeed normally distributed (e.g., doping concentration has a normal distribution), the dependence of the electrical device parameters and gate delay on these physical parameters may not be linear, giving rise to non-normal gate delays. With reduction of geometries, process variation is becoming more pronounced, and the linear approximation may not be accurate for some parameters.

Typically, there are two types of SSTA techniques: Monte Carlo methods and probabilistic analysis methods. In contrast to Monte Carlo based methods, which are based on sample-space enumeration, probabilistic methods explicitly model gate delay and arrival times with random variations. These methods typically propagate arrival times through the timing graph by performing statistical sum and maximum operations. They can be classified into two broad classes: path-based approaches and block-based approaches. In path-based SSTA algorithms, a set of paths, which is likely to become critical, is identified, and a statistical analysis is performed over these paths to approximate the circuit-delay distribution. The basic advantage of this approach is that the analysis is split into two parts—the computation of path delays followed by the statistical maximum operation over these path delays. However, the difficulty with the approach is how to rigorously find a subset of candidate paths such that no path that has significant probability of being critical in the parameter space is excluded. In addition, for balanced circuits, the number of paths that must be considered can be very high. On the other hand, the block-based methods follow the DSTA algorithm more closely and traverse the circuit graph in a topological manner.

In both block-based and path-based SSTA approaches, the gate timing models play a significant role for the accuracy-efficiency trade-off. In function-based SSTA the gate delay is modeled as a linear or non-linear function [29] of process variations, similar to the traditional non-linear delay model [30] in STA. The coefficients are characterized and stored in look-up tables with input slew (S_{in}) and load effective capacitance (C_{eff}) as parameters. When calculating statistical gate delay moments, these coefficients are interpolated based on the nominal value of S_{in} and C_{eff}. However, due to process variations, both S_{in} and C_{eff} are variational as well. Not considering the statistical nature of S_{in} and C_{eff} can result in 30 % delay errors [31]. Also, similar to non-linear delay model, function-based models do not account for resistive interconnect loads and nonlinear input waveforms. Additionally, the function-based delay representation is entirely based on non-physical or empirical models, which is their major source of inaccuracy [32].

A large number of more physical gate timing models have been proposed for accurate STA, such as voltage-dependent current source models [31–39] and transistor-level gate models [40–48]. These gate timing models, denoted as voltage-input voltage-output gate models, represent every gate by current sources and capacitances with respect to input voltage (Vi) and output voltage (Vo). Most voltage-dependent current source models target only accurate modeling of combinational gate delay with the assumptions of single input switching and that the input signal is independent of the output signal. Hence, they fail to model internal nodes and capacitances, which lead to different undesired symptoms for sequential elements, including non-monotonic behavior, failure to model storage behavior, etc. [37]. In contrast, the transistor-level gate models can handle sequential circuits in the same way as the combinational circuits without the limiting assumptions of current source models and are able to consider multiple input (near-)simultaneous switching (MISS). Additionally, the transistor-level gate models have a better defined physical relationship with node voltages and physical parameters and are

more general and accurate for timing, noise and power analysis and practical for multi-million gate STA runs [40–48]. The transistor-level gate models are utilized to estimate the timing variabilities based on corner-based timing analysis in [44]. However, these methods do not take signal correlations and sequential cells into consideration, and most of them are just verified in several simple single gates considering only single input switching. Additionally, the solvers proposed for these statistical delay calculations either have difficulties for other gate timing models [32, 33] or require many simulation trails [31, 34, 44].

In this section, we present a novel method to extend voltage-based gate models for statistical timing analysis. Correlations among input signals and between input signal and delay are preserved during simulation by using same model format for the voltage and all elements in gate models. In the statistical solver, all input signals and their correlations are considered together, thus fundamentally addressing MISS in statistical timing analysis. The variational waveform for statistical delay calculation is computed with random differential equation-based method. For high accuracy in the case of large process variations, the statistical solver divides the process variation space into several sub-spaces and performs the statistical timing analysis in each sub-space. Since a common format for voltage and current waveforms and passive components (resistances and capacitances) is utilized in the gate models, the correlations among input signals and between input signal and delay are preserved during statistical delay calculation. Furthermore, since described timing analysis is based on the transistor-level gate models, it is able to handle both combinational and sequential circuits.

2.3.1 Statistical Simplified Transistor Model

In transistor-level gate models [40–45, 47, 48], the transistor model needs to capture sufficient second-order effects for accuracy, accounting for the impact of process variations, while still being simple enough to be evaluated efficiently. The transistor model for timing analysis in [48] uses look-up tables for drain-source current and an input-transition dependent constant value for five intrinsic capacitances of each transistor. The look-up table based transistor models in [41, 44, 45] implement SPICE's model version for the five intrinsic capacitances. If linear-centric method is utilized, in which the Jacobian matrix is constant for all iterations, the efficiency of transistor-level timing analysis is significantly improved [41, 44, 45, 47]. Current source models require transient analysis or ac analysis for different combinations of S_{in} and C_{eff} or different combinations of input and output voltages at different corners. For transistor-level gate modeling, only characterization of the unique transistors in the standard cell library is needed. The current and capacitances of SSTM are obtained by a dc sweep at the gate, drain and source terminals. For statistical analysis, the sensitivities in SSTM are characterized by a finite-difference approximation.

CMOS transistor drain current modeling: Generally, the MOS transistor drain current I_{DS} is modeled by compact models like BSIM4. With several hundred process parameters, BSIM3/4 determines drain current and sixteen intrinsic capacitances by solving complex equations, which are functions of the process parameters in the model. The physical properties are accurately represented by those parameters, however, the huge amount of computation time makes it impractical for fast timing analysis.

Avoiding approximating data to expressions, the model described in this section addresses these issues by directly using measured or simulated data. Moreover, in comparison with advanced analytical models, this table-based model gains significant speed advantage by using the efficient interpolation and extrapolation methods and resourceful implementation of look-up table sizes.

In nanometer technology, V_T is not only a function of V_{BS} but also V_{DS}, which implies that a 2D look-up table for I_{DS} with entries V_{DS} and $V_{GS} - V_T$ is not practical. The $I_{DS}(V_{GS},V_{DS})$ characteristics have almost the same shape under different V_{BS} when V_{BS} is not close to the supply voltage, implying a possibility of reducing data points corresponding to V_{BS}. For constant V_{BS}, I_{DS} displays different nonlinearity in three operating regions. In the linear region, the current I_{DS} increase rapidly along with V_{DS} while shows nearly linear dependence on V_{DS} with relatively much slower slope in the saturation region. In the cutoff region, however, the current is close to zero and shows a weak relationship with V_{DS} and V_{GS}. In [49], a continuous piecewise linear surface is generated for the current curve using trilinear interpolation [50], mainly due to its reduced complexity in comparison with explicit model evaluation and monotonic piecewise cubic interpolation [51] or spline cubic Hermit interpolation [52]. If derivative of the current is not continuous, Broyden's method [49] avoids the derivative calculation at every iteration by replacing it with finite difference approximation.

Transistor capacitance modeling: The transient response of a combinational logic gate is sensitive to the transistor intrinsic capacitances in the gate. If the intrinsic capacitances are not modeled accurately, the error introduced can accumulate when the transient pulse propagates through the logic chain. Gate level models model a gate capacitance to a constant value C_{eff} ignoring the nonlinear property of the intrinsic capacitances hidden in the gate. One way to model nonlinear intrinsic capacitances is to represent them as voltage-dependent terminal charge sources as in BSIM4. The sixteen capacitances of a transistor are computed from the charge Q by $C_{ij} = \partial Q_i/\partial V_j$ at every time step, where i and j denote the transistor terminals. Although this method may be the most accurate by means of sophisticated charge formulations, the performance and characterization runtime poses the complexity challenges for S/STA.

In the 45 nm node and beyond, the intrinsic capacitance becomes increasingly nonlinear. In order to accurately capture the capacitances, analytical models still play a dominant role in transistor-level timing analysis [44, 45, 50, 53–55]. In [48], the constant capacitance values based on the initial state (cutoff or linear state) are used for the entire transition. However, the assumption that the capacitances influence the output waveform mostly at the beginning would result in deviations

at the end of the transition, adding errors for output slew due to the strong capacitance nonlinearity. In order to improve accuracy while still maintaining satisfactory computational efficiency, the model in [49] treats the five capacitances differently. The gate capacitances C_{GS}, C_{GD} and C_{GB} use 2D look-up tables (as a function of V_{GS} and V_{DS}), while constant values are characterized for junction capacitances C_{SB} and C_{DB}. C_{SB} is at least one order of magnitude smaller than the other capacitances, and normally, C_{DB} is negligible compared to output load. As a consequence, using constant values for C_{SB} and C_{DB} promises fast performance without accuracy loss.

Statistical extension: In addition to the nominal values for the dc current source and intrinsic capacitances, the statistical extension of the model contains the sensitivities of these model elements to any statistical parameter of interest.

The statistical description of the current and the intrinsic capacitance in the model are evaluated as $I_{DS}(\Delta_p) = I_{DS,nom} + \delta I_{DS}(\Delta_p)$ and $C_j(\Delta_p) = C_{j0} + \delta C_j(\Delta_p)$, where p is the random parameter, which is the sum of nominal value $p_{k\rho}$ and random variable η with zero mean μ and standard deviation σ. These process parameters can be physical process parameters such as effective channel length L_{eff}, and threshold voltage V_T, or non-physical parameters derived from dimension-reduction methods, such as principal component analysis, independent component analysis [56, 57], and reduced rank reduction [58]. Δ_p is the parameter deviation from the nominal value p_0 sampled from η and $C_{j\rho}$ is the nominal value of the *jth* capacitance. Note that the correlations among the statistical variables are submissive to accuracy-speed trade-off. The numerical sensitivity is characterized by perturbing the statistical parameter being modeled above and below (e.g. $\pm\sigma$) its nominal value. Since nowadays standard cell libraries consist of hundreds of cells with many process corners, gate level models require a significant amount of *cpu* time to characterize all the standard cells. The described transistor-level gate model has modest characterization requirements: it only needs to characterize the unique transistors in the cell library. It is also worth mentioning that I_{DS} and the gate capacitances are roughly proportional to *W*/*L* and *WL*, respectively, raising the possibility to require only a few table models for each MOST type.

2.3.2 Bounds on Statistical Delay

The process variation vector η includes both global process variations and local variations. For a specific random process parameter with a global deviation and local deviations, the global deviation and correlated local deviation affect all the transistors in the same way hence they can be clubbed together [59]. The large number of local process deviations can be significantly reduced to a much smaller number of independent local variables with techniques like principal component analysis. According to [48, 59], the local variables can be further collapsed to a single variable by treating it as in a root of the sum of square technique. For voltage-input voltage-output gate models, like current source models and

transistor-level gate models in [31–34, 40–48], nodal analysis or modified nodal analysis is used for gate simulation. Rewriting (2.7) as

$$F(q', q, t, \eta) = 0 \tag{2.14}$$

the first-order Taylor piecewise-linearization of (2.14) in x_0 yields

$$P(x_0)\xi' = K(x_0)\xi + L(x_0)\eta \tag{2.15}$$

where P, K and L are matrices defined as $\partial F/\partial x'_0$, $\partial F/\partial x_0$, $\partial F/\partial p$, respectively. Transient analysis requires only the solution of the deterministic version of (2.14), e.g. by means of a conventional circuit simulator, and of (2.15) with a method capable of dealing with linear stochastic differential equations with stochasticity that enters only through the initial conditions. Since (2.15) is a linear homogeneous equation in ξ, its solution, will always be proportional to $\eta - \eta_0$. According to [60], (2.8) has a unique mean square solution which can be represented by $\xi(t) = \Gamma(t)(\eta - \eta_0)$. Following the procedure as described in Sect. 2.2, (2.15) for $\Gamma(t)$ can be written as

$$P(x_0)\Gamma'(t) = K(x_0)\Gamma(t) + L(x_0) \tag{2.16}$$

In delay distribution calculation, at every time point, P, K and L are updated and function (2.16) can be solved to obtain $\Gamma(t)$. If $\Gamma(t)$ and L have high dimension (e.g. number of process variations is large), the sensitivity of the variational voltage to the jth process variation, must be computed. Based on (2.16), $\Gamma_j(t)$ is calculated as

$$P(x_0)\Gamma'_j(t) = K(x_0)\Gamma_j(t) + L(x_0)u \quad j = 1 : p_m \tag{2.17}$$

where u is selection vector whose elements are all zeros except the jth element, which has value one. After using a numerical integration method, due to x_o-dependent coefficients $P(x_0)$, $K(x_0)$ and $L(x_0)$, (2.17) becomes a linear algebraic equation with respect to the variable $\Gamma_j(t)$. The covariance matrix (2.11) of the solution, rewritten here for clarity, is expressed as

$$\Sigma_{\xi\xi} = \Gamma\Sigma_{\eta\eta}\Gamma^T \tag{2.18}$$

To extend voltage-input voltage-output gate models for statistical timing analysis, in addition to statistical simulation, the extraction of statistical delay from variational voltages is also necessity. The extraction methods of existing gate level statistical timing analysis have the three main categories: interpolation-based analysis, Monte Carlo simulation based on statistical current source models and direct calculation based on Markovian process assumption. In interpolation-based analysis [44] the output waveforms at different corners are simulated, and then the output waveform is characterized by linear interpolation. However, this method assumes that the results at different corners are linear with respect to the process variations and large number of samples is required for delay calculation. The statistical moments of several crossing times are calculated by Monte Carlo

simulations based on statistical current source models in [31, 34]. However, even though the Monte Carlo simulations are applied, the accuracy of statistical delay calculation is not competitive due to the over-simplified current source models. In direct calculation based on Markovian process assumption, the delay distribution is calculated by assuming that the voltage at every time point is a Markovian stochastic process due to the numerical integration method [32, 61, 62]. In order to calculate the distribution of a crossing time, the joint probability of voltage at different time steps is calculated by using the bivariate normal distribution, which is erroneous when the Gaussian distribution assumption for voltages is inaccurate. Here, the boundaries of voltage of interest, which needs to be stored and propagated (denoted as Ξ_r with mean value $\mu_{\Xi r}$)), can be expressed as

$$[\Xi_{r,\min}, \Xi_{r,\max}] = \mu_{\Xi_r} \pm \sum_k \sum_m \{|\Sigma_{\xi\xi}|^{\max}\} \tag{2.19}$$

for any $p_i \in \{p_1,\ldots,p_m\}$ of $i \in \{i_1,\ldots,i_k\}$ transistors connected to node $r \in \{r_1,\ldots,r_q\}$. In this scheme higher order moments are expressed in terms of the first and second order moments as if the components of Ξ_r are Gaussian processes. The method is fast, and comparable to regular nominal circuit simulation. Suppose that there are m-trial Monte Carlo simulation for n faults, the method (using statistical data of the process parameters variations) gains a theoretical speed-up of $m \times n$ over the Monte Carlo method. During path-based timing analysis, each critical path can be simulated as a whole to obtain $\mu_{\Xi r}$ and Ξ_r directly for statistical path delay calculation. Gate-by-gate propagation can also be used. For a single transition propagating from gate to gate, $\mu_{\Xi r}$ and Ξ_r of each gate during the transition period (when $\mu_{\Xi r}$ switches from low to high or from high to low) are propagated. This expresses the voltages as linear functions of the process variables, through which the correlations between voltages are implicitly defined. During statistical timing analysis, the correlation of signals caused by process variations and path re-convergence should be considered and efficiently simulated.

Here, if more than one input switch in a multi-input gate, the 50 % crossing time standard deviation σ of every two switching inputs are calculated and checked. If the signals are not overlapping, the correlation between them will be ignored and the latest/earliest input or inputs will be propagated while the other is assumed static. On the other hand, if they are overlapping, all stochastic correlated inputs are considered.

2.3.3 Reducing Computational Complexity

The gate models are constructed by replacing every transistor in the gate by its corresponding SSTM. After *RC* extraction, model order reduction (MOR) techniques is employed to reduce the complexity of the interconnect model, in which every resistance and capacitance is represented as a linear function of process

variations. In an asymptotic waveform evaluation (AWE) algorithm [63] explicit moment matching was used to compute the dominant poles via Padé approximation. As the AWE method is numerically unstable for higher-order moment approximation, a more elegant solution to the numerical problem of AWE is to use projection-based MOR methods. In the Padé via Lanczos (PVL) method [64], the Lanczos process, which is a numerically stable method for computing eigenvalues of a matrix, was used to compute the Krylov subspace. In PRIMA [65] the Krylov subspace vectors are used to form the projector for the congruence transformation, which leads to passive models with the matched moments in the rational approximation paradigm. However, these methods are not efficient for circuits with many inputs and output terminals as the reducing cost are tied to the number of terminals; the number of poles of reduced models is also proportional to the number of terminals. Additionally, PRIMA-like methods do not preserve structure properties like reciprocity of a network.

Another approach to circuit-complexity reduction is to reduce the number of nodes in the circuits and approximate the newly added elements in the circuit matrix in reduced rational forms by approximate Gaussian elimination for *RC* circuits [66]. Alternatively, model order reduction can be performed by means of singular-value-decomposition (SVD) based approaches such as control-theoretical-based truncated balance realization (TBR) methods, where the weakly uncontrollable and unobservable state variables are truncated to achieve the reduced models [67–73]. The major advantage of SVD-based approaches over Krylov subspace methods lies in their ability to ensure the errors satisfying an a priori upper bound [71]. Also, SVD-based methods typically lead to optimal or near optimal reduction results as the errors are controlled in a global way, although, for large scale problems, iterative methods have to be used to find an adequate balanced approximation (truncation). In this respect, ideas based on balanced reduction methods are significant since they offer the possibility to perform order selection during the computation of the projection spaces and not in advance. Typically in balanced reduction methods, there is a rapid decay in the Gramians eigenvalues. As a consequence these Gramians can be well approximated using low-rank approximations, which are used instead of the original. Accordingly, several SVD approaches approximate the dominant Cholesky factors (dominant eigensubspaces) of controllability and observability Gramians [68, 72, 73] to compute the reduced model.

In this section, we adjust the dominant subspaces projection model reduction (DSPMR) [68] and provide an approximate balancing transformation for circuits whose coefficient matrices are large and sparse such as in interconnect. The approach presented here produces orthogonal basis sets for the dominant singular subspace of the controllability and observability Gramians significantly reducing the complexity and computational costs of singular value decomposition, while preserving model order reduction accuracy and the quality of the approximations of the TBR procedure.

In the analysis of delay or noise in on-chip interconnect we study the propagation of signals in the wires that connect logic gates. These wires may have

numerous features: bends, crossings, vias, etc., and are modeled by circuit extractors in terms of a large number of connected circuit elements: capacitors, resistors and more recently inductors. Given a state-space formulation of the interconnect model

$$\begin{aligned} C(dx/dt) &= Gx(t) + Bu(t) \\ y(t) &= E^T x(t) \end{aligned} \tag{2.20}$$

where $C, G \in R^{n\times n}$ are matrices describing the reactive and dissipative parts of the interconnect, respectively, $B \in R^{n\times p}$ is a matrix that defines the input ports, $E \in R^{p\times n}$ is matrix that defines the outputs, and $y(t) \in R^q$ and $u(t) \in R^p$, are the vectors of outputs and inputs, respectively, the model reduction algorithm seek to produce a similar system

$$\begin{aligned} \widehat{C}d\widehat{x}/dt &= \widehat{G}\widehat{x}(t) + \widehat{B}u(t) \\ \widehat{y}(t) &= \widehat{E}^T \widehat{x}(t) \end{aligned} \tag{2.21}$$

where $\widehat{C}, \widehat{G} \in R^{k\times k}$, $\widehat{B} \in R^{k\times m}$, $\widehat{E} \in R^{p\times k}$, of order k much smaller than the original order n, but for which the outputs $y(t)$ and $\widehat{y}(t)$ are approximately equal for inputs $u(t)$ of interest. The Laplace transforms of the input output transfer functions

$$\begin{aligned} H(s) &= E^T(G + sC)^{-1}B \\ \widehat{H}(s) &= \widehat{E}^T(\widehat{G} + s\widehat{C})^{-1}\widehat{B} \end{aligned} \tag{2.22}$$

are used as a metric for approximation accuracy if

$$\left\| H(s) - \widehat{H}(s) \right\| < \varepsilon \tag{2.23}$$

for a given allowable error ε and an allowed domain of the complex frequency variable s, the reduced model is accepted as accurate.

Balanced truncation [67, 73], singular perturbation approximation [74], and frequency weighted balanced truncation [75] are model reduction methods for stable systems. Except for modal truncation each of the above methods is based either explicitly or implicitly on balanced realizations, the computation of which involves the solutions of Lyapunov equations

$$\begin{aligned} GXC^T + CXG^T &= -BB^T \\ G^T YC + C^T YG &= -E^T E \end{aligned} \tag{2.24}$$

where the solution matrices X and Y are controllability and observability Gramians. The original implementation of balanced truncation [67] involves the explicit balancing of the realization (2.20). This procedure is dangerous from the numerical point of view because the balancing transformation matrix T tends to be highly ill-conditioned. The square root method [73] is an attempt to cope with this problem

by avoiding explicit balancing of the system. The method is based on the Cholesky factors of the Gramians instead of the Gramians themselves. In [76] the use of the Hammarling method was proposed to compute these factors. Recently, in [68] and [72] it has been observed that solutions to Lyapunov equations often have low numerical rank, which means that there is a rapid decay in the eigenvalues of the Gramians.

Indeed, the idea of low-rank methods is to take advantage of this low-rank structure to obtain approximate solutions in a low-rank factored form. The principal outcome of these approaches is that the complexity and the storage are reduced from $O(N^3)$ flops and $O(N^2)$ words of memory to $O(N^2r)$ flops and $O(Nr)$ words of memory, respectively, where r is the approximate rank of the Gramian ($r \ll N$). Moreover, approximating the Cholesky factors of the Gramians directly and using these approximations to provide a reduced model, has a comparable cost to that of the popular moment matching methods. It requires only matrix-vector products and linear solvers.

For large systems with a structured transition matrix, this method is an attractive alternative because the Hammarling method can generally not benefit from such structures. In the original implementation this step is the computation of exact Cholesky factors, which may have full rank. We formally replace these (exact) factors by (approximating) low rank Cholesky factors [68, 72]. The iterative procedure approximates the low rank Cholesky factors Z_X and Z_Y with r_X, $r_Y \ll n$, such that $Z_X Z_X^H \approx X$ and $Z_Y Z_Y^H \approx Y$, where H is Hermitian (complex-conjugate) matrix. Note that the number of iteration steps i_{max} needs not be fixed a priori. However, if the Lyapunov equation should be solved as accurate as possible, correct results are usually achieved for low values of stopping criteria that are slightly larger than the machine precision. Let

$$Z_Y^H Z_X = U_Y \Sigma U_X^H \tag{2.25}$$

be SVD of $Z_Y^H Z_X$ of dimension $N \times m$. The cost of this decomposition including the construction of U is $14Nm^2 + O(m^3)$ [77]. To avoid this, we perform eigenvalue decomposition

$$(Z_Y^H Z_X)^H Z_Y^H Z_X = U_Y \Lambda U_X^H \tag{2.26}$$

Comparing (2.26) with (2.25) shows that the same matrix U_X is constructed and that

$$(Z_Y^H Z_X U_X)^H Z_Y^H Z_X U_Y = \Lambda = \Sigma^H \Sigma \tag{2.27}$$

This algorithm requires Nm^2 operations to construct $(Z_Y^H Z_X)^H Z_Y^H Z_X$ and $Nmn + O(m^3)$ operations to obtain $Z_Y^H Z_X U_X \Sigma^{-1}$ for $n \times n$ Σ. The balancing transformation matrix T is used to define the matrices $S_X = T_{(1:k)}$ and $S_Y = T_{(1:k)}^{-T}$. If $\sigma_k \neq \sigma_{k+1}$, the reduced order realization is minimal, stable, and balanced, and its Gramians are equal to $diag(\sigma_1,\ldots,\sigma_k)$. The balancing transformation matrix can be obtained as

$$S_X = Z_X U_X \Sigma^{-1/2} \quad S_Y = Z_Y U_Y \Sigma^{-1/2} \tag{2.28}$$

then, under a similarity transformation of the state-space model, both parts can be treated simultaneously after a transformation of the system (C, G, B, E) with a nonsingular matrix $T \in R^{n \times n}$ into a balanced system

$$\widehat{C} = S_X C S_Y^H \quad \widehat{G} = S_X G S_Y^H \quad \widehat{B} = S_Y^H B \quad \widehat{E} = E S_X \tag{2.29}$$

In this algorithm we assume that $k \leq r$ ($rank\ Z_Y^H Z_X$). Note that SVDs are arranged so that the diagonal matrix containing the singular values has the same dimensions as the factorized matrix and the singular values appear in non-increasing order.

2.4 Yield Constrained Energy Optimization

One of the most notable features of ultra-low power nanometer-scale CMOS circuits is the increased sensitivity of circuit performance to process parameter variation when operating at reduced V_{DD} supplies. The growth of variability can be attributed to multiple factors, including the difficulty of manufacturing control, the emergence of new systematic variation-generating mechanisms, and most importantly, the increase in fundamental atomic-scale randomness, such as the variation in the number of dopants in the transistor channel [5]. As a consequence, device upsizing may be required to achieve operational robustness against process variability at the expense of a higher energy consumption and larger area occupation [6]. Technology scaling, circuit topologies, and architecture trends have all aligned to specifically target low-power trade-offs through the use of fine-grained parallelism [7], near-threshold design [8], V_{DD} scaling and body biasing [9]. Similarly, a cross-layer optimization strategy is devised for variation resilience, a strategy that spans from the lowest level of process and device engineering to the upper level of system architecture. As a result, power-management has evolved from static custom-hardware optimization to highly dynamic run-time monitoring, assessing, and adapting of hardware performance and energy with precise awareness of the instantaneous application demands. These mechanisms allow to dynamically select the most appropriate operating point for a particular process corner that affects the die and its sub-components. Simultaneous circuit yield and energy optimization with key parameters (supply voltage V_{DD} and supply to threshold voltage ratio V_{DD}/V_T) is a part of a system-wide strategy, where critical parameters that minimize energy (e.g. V_{DD}/V_T) provide control mechanisms (e.g. adaptive voltage scaling) to run-time system. Yield constrained energy optimization, as an active design strategy to counteract process variation in sub-threshold or near-threshold operation, necessitates the need for statistical design paradigm to overcome the limitations of deterministic optimization schemes, such as sizing [78] and dual-V_T allocation [79]. Analytical optimization based on sensitivities

[80], fitted [81] and physical [82] parameters offer guidelines for optimum power operation. The choice of the nonlinear optimization techniques [83–85] is based on the nonlinear relationships that exist between device lengths and widths and their associated delays, particularly with strong short-channel effects in the nanometer region, and leakage power.

In this section, we extend nonlinear optimization by developing a yield constrained sequential energy minimization framework that is applied to multivariable optimization in body bias enabled subthreshold and near-threshold designs. The presence of the yield constraint in nonlinear optimization makes the problem non-convex, thus hard to solve in general. In the proposed algorithm, we create a sequence of minimizations of the feasible region with iteratively-generated low-dimensional subspaces. As the resulting sub-problems are small, global optimization in both convex and non-convex cases is possible. The method can be used with any variability model, and is not restricted to any particular performance constraint. The yield constraint becomes active as the optimization concludes, eliminating the problem of overdesign in worst-case approach.

2.4.1 Optimum Energy Point

The optimum energy point arises from opposing trends in the dynamic and the leakage energy consumed per clock cycle as supply voltage V_{DD} scales down. The dynamic (CV^2) energy decreases quadratically, but in the subthreshold region, the leakage energy per cycle increases as a result of the leakage energy being integrated over exponentially longer clock periods. With process scaling, the shrinking of feature sizes implies smaller switching capacitances and thus lower dynamic energy consumed. At the same time, leakage current in recent technology generations have increased substantially, in part due to threshold voltage V_T being decreased to maintain performance while the nominal supply voltage is scaled down. On a chip-level, energy consumption is optimized by adjusting V_{DD} (dynamic supply voltage scaling) and V_T (body-biasing) within its functional operating region (defined by its local process variations, i.e. the distributions of the critical dimension size, oxide thickness, and threshold voltage). The mean value of the performance range at a particular temperature or voltage is determined by the semiconductor process corner—an aggregation of process variations effects—that impacts the circuit. The range width is determined by process, voltage and temperature variations, which impose V_{DD} to V_T ratio, noise margins and thus limit the performance range. Consider the delay d_j of path j,

$$d_j = V_{DD} \sum_{i \in j} (C_{intr,i} + x_i^{-1} C_{extr,i}) I_{drive,i}^{-1} e^{k_i V_{BB}} \leq T_{clk} \quad \forall j \in \Lambda \tag{2.30}$$

where i is an index that runs over all gates in the circuit, j is an index that runs over all circuit paths, Λ is the collection of all paths in the circuit, x is the gate sizing

factor ($x \geq 1$), C_{intr} and C_{extr} are the switching intrinsic and extrinsic capacitance of a gate, respectively, I_{drive} is the current drive of a gate, V_{BB} represents the symmetrical forward body-bias voltage ($V_{BB} = V_{DD} - V_{nwell} = V_{pwell}$), T_{clk} is the operating clock period and k is fitting parameter. Expression (2.30) constrains the delay of each circuit path to be less than the targeted clock period, T_{clk}. The dependence of $C_{intr,i}$ on body-bias is accounted for through fitting parameter k_i. Based on the above model, the total energy of a CMOS digital circuit design under body-bias conditions is modeled as [86]

$$E_{total} = V_{DD} \sum_{i=1}^{N} \left(\begin{array}{c} a\left(\frac{x_i C_{intr,i}}{(1 - m_1 V_{BB})^{m_2}} C_{extr,i} \right) V_{DD} \\ + T_{ck} x_i I_{leak,i} \left(e^{l_{1i} V_{BB}} + l_{2i} \left(e^{l_{1i} V_{BB}} - 1 \right) \right) \end{array} \right) \forall V_{BB} \geq 0 \qquad (2.31)$$

where a is the average circuit activity factor, N is the total number of gates in the circuits, and l_1, l_2, l_3, m_1 and m_2 are fitting parameters. At a given V_{DD}, the lowest energy design is obtained when no gates are up-sized, e.g. $x_i = 1 \ \forall$ gates i. However, this also leads to the slowest design, as can be inferred from (2.30). We model the manufactured values of the parameters $p_k \in \{p_1, \ldots, p_m\}$ for transistor k as a random variable

$$p_k = \mu_{p,k} + \sigma_p(\lambda_k) \cdot p(\lambda_k, \theta) \qquad (2.32)$$

where $\mu_{p,k}$ and $\sigma_p(\lambda_k)$ are the mean value and standard deviation of the parameter p_k, (e.g. channel-length L, threshold voltage V_T) respectively, $p(\lambda_k,\theta)$ is the stochastic process corresponding to parameter p, λ_k denotes the location of transistor k on the die with respect to a point origin and θ is the die on which the transistor lies. Assuming that p_k is a zero-mean Gaussian process and using the Karhunen-Loève expansion, p_k can be written in truncated form (for practical implementation) by a finite number of terms Ψ as in Sect. 2.1 [87]

$$p_k = \mu_{p,k} + \sigma_p(\lambda_k) \cdot \sum_{n=1}^{\Psi} \sqrt{\vartheta_{p,n}} \delta_{p,n}(\theta) f_{p,n}(\lambda_k) \qquad (2.33)$$

where $\{\delta_n(\theta)\}$ is a vector of zero-mean uncorrelated Gaussian random variables and $f_{p,n}(\lambda_k)$ and $\vartheta_{p,\,n}$ are the eigenfunctions and the eigenvalues of the covariance matrix $\Sigma_p(\lambda_1, \lambda_2)$ of $p(\lambda_k, \theta)$, controlled through a distance based weight term, the measurement correction factor, correlation parameter ρ and process correction factors c_x and c_y.

The optimization problem, given r iterations, is than formulated as to find a design point d^* that minimizes total energy E_{total} over design variable vector d (e.g. gate size W, supply voltage V_{DD}, bulk-to-source voltage V_{BS}, etc.) in the design space Φ, subject to a minimum delay d_j of path j and a minimum yield requirement y given bound β

$$
\begin{aligned}
&d'^{*} = \underset{d' \in \Phi(E_{total})}{\arg\min}\, E_{total}(d') \\
&\text{subject to} \\
&y_r(d'_r) = \mathrm{EV}\{y_r(d'_r, p^v_{k,r}) | pdf(p^v_{k,r})\} \\
&y_r(d'_r, p^v_{k,r}) \geq 1 - \beta \quad v = 1, \ldots M \quad \forall d' \in \Phi(E_{total,r}) \\
&d_{j,r} \leq T_{clk} \quad \forall j \in \Lambda \\
&x_i = 1 \quad \forall i \in \{1, 2, \ldots, q\}
\end{aligned}
\tag{2.34}
$$

where EV is the expected value and each vector d' has an upper and lower bound determined by the technological process variation p with probability density function $pdf(d)$ and $p^1,\ldots,p^M$ are M (independent) realizations of the random vector p. Let $\Phi(E_{total})$ be the compact set of all valid design variable vectors d' such that $E_{total}(d) = E_{total}$. That Φ is assumed to be compact is, for all practical purposes, no real restriction when the problem has a finite minimum. The main advantage of this approach is its generality: it imposes no restrictions on the distribution of p and on how the data enters the constraints. If, as an approximation, we restrict $\Phi(E_{total,r})$ to just the one-best derivation of $E_{total,r}$, then we obtain the structured perceptron algorithm [88]. As a consequence, given active constraints including optimum energy budget and minimum frequency of operation, (2.34) can be effectively solved by a sequence of minimizations of the feasible region with iteratively-generated low-dimensional subspaces.

2.4.2 Optimization Problem

To start the optimization problem, a design metric for global solution is initially selected, based on the priority given to the energy budget as opposed to the performance function in a given application. In the algorithm, we use a cutting plane method [89] to repeatedly recomputed optimum design point d'^{*} with a precision of at least ε and add it to a working set S_r of derivations on which (2.34) is optimized. A new d'^{*} is added to the working set only if $d'^{*} > \varepsilon$; otherwise, the algorithm terminates, e.g. we are cutting out the halfspace because we know that all such points have an objective value larger than ε, hence can not be optimal. The algorithm solves (2.34) restricted to S_r by sequential minimal optimization [90], in which we repeatedly select a pair of derivatives of d' and optimize their dual (Lagrange) variables, required to find the local maxima and minima of the performance function. Although sequential minimal optimization algorithm is guaranteed to converge, we used the heuristics suggested by [91] to accelerate the rate of convergence and to select feasibility region: one must violate one of the conditions, and the other must allow the objective to be improved. At the end of sequence, we average all the weight vectors obtained at each iteration, just as in the averaged perceptron. The result of this optimization is the minimum energy

design that meets a targeted performance under yield constrains and scaled supply voltage and body bias conditions.

Parameter update: To insure that the data is completely separable, we employ stochastic steepest gradient descent method to adapt the parameters. We map design variable vector d' to feature vectors $h(d')$, together with a vector of feature weights w, which defines contribution of design variable in obtained yield. Updating feature weights is presented as a quadratic program

$$\begin{aligned} &\text{minimize } 1/2\eta\|w' - w\|^2 \\ &\text{subject to } y_r(w, d', p^v_{k,r}) \geq 1 - \beta, \quad v = 1, \ldots M \quad \forall d' \in \Phi(E_{total,r}) \end{aligned} \tag{2.35}$$

where η is a step size. The quadratic programming problem is solved incrementally, covering all the subsets of classes constructing the optimal separating hyperplane for the full data set. If no hyperplane can be found that can divide the a priori and a posteriori classes, with the modified maximum margin technique [92] we find a hyperplane that separates the training set with a minimal number of errors.

Actual risk and optimal bound: The approximation-based approach to processing statistical yield constrained problems requires mechanisms for measuring the actual risk (reliability) associated with the resulting solution, and bounding the true optimal value of the yield constraint problem (2.34). A straightforward way to measure the actual risk of a given candidate solution is to use Monte Carlo sampling. We define a reliable bound on $pdf(d)$ as the random quantity

$$\beta := \arg\max_{\gamma \in [0,1]} \{\gamma : \sum_{s=0}^{\Delta} \binom{M}{s} \gamma^s (1-\gamma)^{M-s} \geq \delta\} \tag{2.36}$$

where 1-δ is the required confidence level. Given candidate solution $d' \in \Phi(E_{total,i})$, the probability $pdf(d')$ is estimated as Δ/M, where Δ is the number of times the condition is violated. Since the outlined procedure involves only the calculation of quantities y_r, it can be performed with a large sample size M, and hence feasibility of d' can be evaluated with a high reliability, provided that β is within realistic assumption.

2.5 Experimental Results

The experiments were executed on a 64-bit Linux server with two quadcore Intel Xeon 2.5 GHz CPUs and 16 GB main memory. The calculation was performed in a numerical computing environment [93]. The effectiveness of the algorithm was evaluated on several circuits exhibiting different distinctive feature in a variety of applications. As one of the representative examples of the results that can be obtained, firstly an application of statistical simulation to the characterization of two analog circuits, the continuous-time bandpass G_m-*C*-*OTA* biquad filter [94]

and discrete time variable gain amplifier is shown. For clarity, the experimental results obtained from these two circuits are illustrated in Sect. 3.5. The statistical timing analysis was characterized by using the BSIM4 model in Spectre and tested on all combinational cells and widely-used sequential cells found in the standard cell library of the Nangate 45 nm open cell library package 2009 [95] and on ISCAS85 benchmark circuits. Spectre can provide the necessary intrinsic capacitance values of each transistor after dc simulation. The Verilog netlists of all ISCAS85 circuits are downloaded from [96] and then mapped to the Nangate 45 nm technology library with Cadence Encounter. The parasitic RC models of the wires are extracted from layout and stored in SPF and SPEF files.

From each circuit the most critical non-false path found by the timing engine in Encounter is extracted. The parser reads the Verilog netlist and SPF files, and then constructs simulation equations for stages, paths and circuits. In order to check the error contributed by the SSTM only, the SSTM model is implemented in Verilog-A and loaded it as a compiled model in Spectre [97].

To characterize the timing behavior, a lookup table-based library is employed which represents the gate delay and output transition time as a function of input arrival time, output capacitive load, and several independent random source of variation for each electrical parameter (i.e., R and C). In each case, both driver and interconnect are included for the stage delay characterizations. The statistical simulation depends on the nominal value computation. As a consequence, firstly the accuracy of the gate models for deterministic timing analysis (no process variations) is evaluated on the minimum-sized standard cells. In the experiments, every switching input signal is a ramp with input slew varying from 7.5 to 600 ps and the load capacitance changes from 0.40 to 25.6 fF. The input slew and load capacitance ranges are the same as the ranges in the non-linear delay model liberty file of the library. Both rising and falling inputs are simulated. Additionally, the scenarios that all input signals switch at the same time are also included. For every gate, hundreds of simulations are performed for different input slew, output capacitance and input switching scenarios, which result in hundreds of delay and slew errors. The average error of the model relative to SpectreB for delay and slew errors is 0.47 and 0.2 % for mean and 0.28 and 0.91 % for standard deviation, respectively. The accuracy of the model and the deterministic simulation method is also evaluated on the critical paths of the ISCAS circuits. The delay and slew errors are within 1 and 2 % of SpectreB indicating high accuracy of the LUT-based simplified transistor model for timing analysis. The statistical simulation method is evaluated also on cells with up to four inputs that have a high probability to switch near-simultaneously. All input signals of these gates are variational with variable correlation.

The variational input signals are modeled as a ramp signal of 40 ps mean input transition time with voltage variations. Two parameters are varied to obtain diverse scenarios to simulate for every cell: the standard deviation of input voltages and nominal arrival time differences between every two input signals. The minimum and maximum of standard deviation of input voltages are 1 and 10 % of V_{DD}, respectively. The correlations among pairs of voltage variations range from

0 to 0.8. The statistical simulation results are compared to 10 k SpectreB Monte Carlo simulations. The mean errors are within 1 % and errors in standard deviation of delay are lower than 6 %. The third-order statistical central moment, skewness has maximum error of approximately 8 %, which occurs when both the standard deviation of input voltages and the correlation coefficient have the largest value. The average mean, standard deviation and skewness errors across critical paths of ISCAS85 circuits are 0.38, 2.30 and 2.87 %, respectively, which for a statistical delay calculation with multiple input switching seem acceptable. Similarly, three different sequential circuits with increasing level of complexity [98] have been evaluated: (i) an active-high transparent latch composed of 16 transistors, a positive-edge triggered D flip-flop composed of 28 transistors and a sequential circuit [98] with in total 90 transistors. For all these circuits, the standard deviation errors are within 2 %. Compared to SpectreB Monte Carlo runs, the evaluated method, achieves 200 times speed-up on average. The speed-up is smaller for larger circuits, showing the benefit of the sparse matrix techniques and efficient data loading techniques employed in Spectre.

The accuracy to estimate the delay moments considering correlation coefficient highly depends on the sensitivity characterization. The sensitivities of current source model element to process variations are characterized based on best mean square error fit and derived from a series of Spice Monte Carlo simulations in [32]. In order to prevent the explosion of LUTs, [31] model the current and capacitance in gate models as a second order Hermite polynomials of process variations. These methods vary all the process variations of interest together for sensitivity characterization, which takes into account the physical correlation of process parameters. However, such characterization exponentially increases simulation time. In the method shown in Sect. 2.3.1, very fast, simple finite differences method is employed for sensitivity approximation (only one or two extra dc analysis are required for each transistor) at the cost of small loss of accuracy.

The analytical delay distribution obtained using the quadratic interconnect model in 45 nm CMOS technology is illustrated in Fig. 2.2a. The nominal value of the total resistance of the load and the total capacitance is chosen from the set 0.15–1 kΩ and 0.4–1.4pF, respectively. The sensitivity of each given data to the sources of variation is chosen randomly, while the total σ variation for each data is chosen in the range of 10–30 % of their nominal value. The scaled distribution of the sources of variation is considered to have a skewness of 0.5, 0.75, and 1. For model order reduction we consider a *RC*-chain with 2002 capacitors and 2003 resistors In Fig. 2.2b, c the convergence history with respect to the number of iteration steps for solving the Lyapunov equation is plotted. For the tolerances at a residual norm of about the same order of magnitude, convergence is obtained after 40 and 45 iterations, respectively. The *cpu*-time needed to solve the Lyapunov equations according to the related tolerance for solving the shifted systems inside the iteration is 2.7 s. Note further that saving iteration steps means that we save large amounts of memory-especially in the case of multiple input and multiple output systems where the factors are growing by p columns in every iteration step. When very accurate Gramians (e.g. low rank approximations to the solutions) are

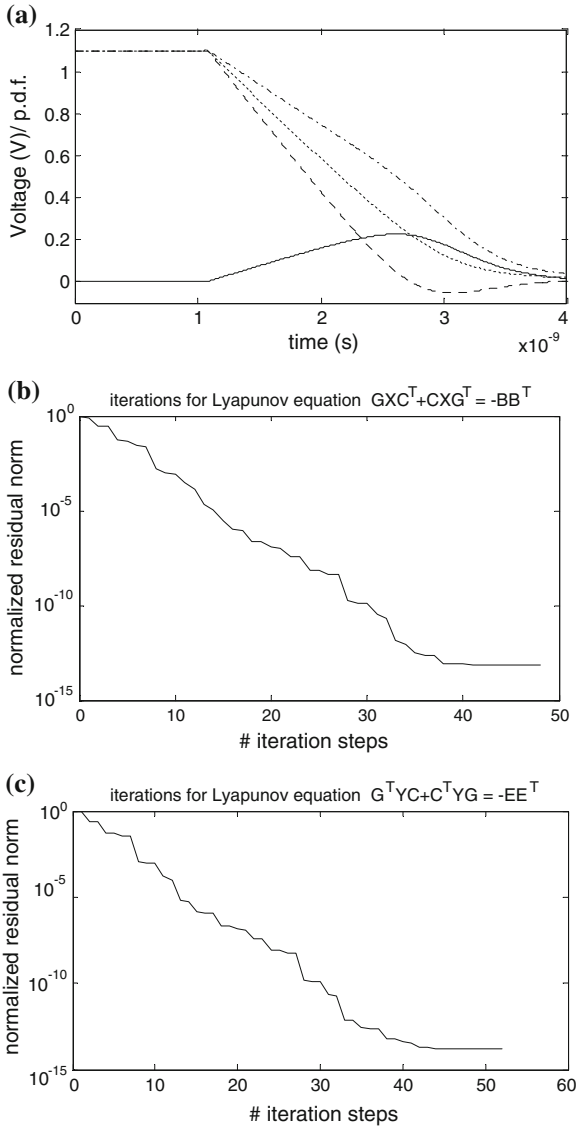

Fig. 2.2 **a** Analytical delay distribution in 45 nm CMOS technology. *Solid line* illustrates delay variance. **b** Convergence history of residual forms. The convergence is obtained after 40 iterations. **c** Convergence history of residual forms. The convergence is obtained after 45 iterations (© IEEE 2011)

selected, the approximation error of reduced system as illustrated in Fig. 2.3a is very small compared to the Bode magnitude function of the original system. The lower two curves correspond to the highly accurate reduced system; the proposed model order reduction technique delivers a system of lower order, and the upper two denote $k = 20$ reduced orders. The frequency response plot is obtained by computing the singular values of the transfer function $H(j\omega)$, which is the frequency response (2.23) evaluated on the imaginary axis (Fig. 2.3b). The error plot is the frequency response plot of the singular values of the error system as a

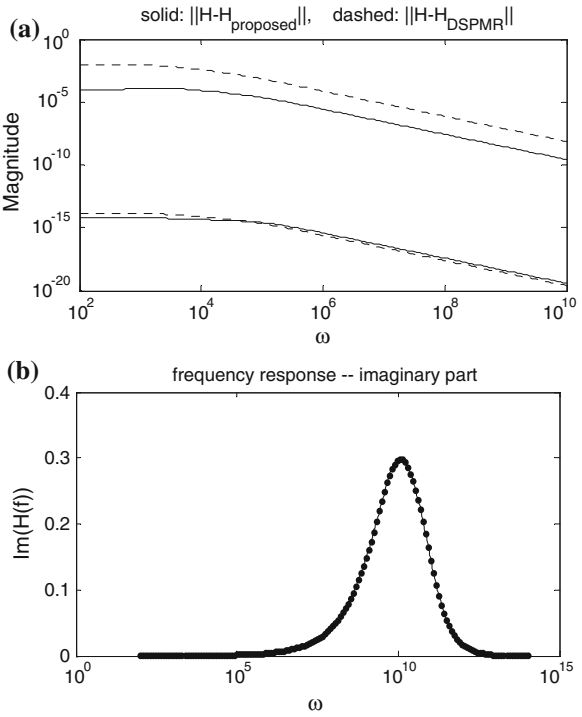

Fig. 2.3 **a** The Bode magnitude plot of the approximation errors. **b** Frequency response of the interconnect model (© IEEE 2011)

function of ω. The reduced order is chosen in dependence of the descending ordered singular values σ_1, σ_2,… σ_r, where r is the rank of factors which approximate the system Gramians. For n variation sources and l reduced parameter sets, the full parameter model requires $O(n^2)$ simulation samples and thus has a $O(n^6)$ fitting cost. On the other hand, the presented parameter reduction technique has a main computational cost attributable to the $O(n + l^2)$ simulations for sample data collection and $O(l^6)$ fitting cost significantly reducing the required sample size and the fitting cost.

To evaluate yield constrained energy optimization BasicMath application from the MiBench benchmark [99] is selected and run on datasets. Switching activities were obtained utilizing SimpleScalar [100]. The calculation was performed in a numerical computing environment [93]. In order to estimate power figures corresponding to execution, the SimpleScalar simulator is used with an online power estimator at different voltage-frequency levels. The constant parameters for the energy and delay models were extracted from HSPICE simulation [101] with UMC 1P8 M 65 nm CMOS model files.

We illustrate the proposed method on a 64-b static Kogge-Stone adder [102] with a 60 μm gate load at its output. The gate-to-gate wire capacitance is included and computed assuming a 4-μm bit pitch. We considered channel-length and threshold-voltage variations with 3 σ/μ of 20 %. These variation levels are

consistent with values in the literature [103]; however, it should be noted that the absolute value of variability is not critical in validating the proposed techniques. All variation in V_T was assumed to be random, due to random-dopant effects.

Energy minimization for fixed input size and fixed output load: As energy consumption becomes more critical, circuit designers are forced to find the globally minimal energy design point for the required delay target under yield constrain. The solution requires the optimization for minimal energy while the delay is fixed. The normalized contours of optimal energy-delay product obtained from energy minimization are shown in Fig. 2.4a. The reference is the design sized for minimum delay under maximum V_{DD} and reference V_T. At this input size, the energy-delay among logic stages is balanced. Therefore, increasing the input size beyond this optimal value will result in more energy consumption. This characteristic of the design, with respect to energy, is distinctive compared to its delay characteristic where the delay is continuously improved by increasing input size. The choice of design region is set by the delay target and the input size condition. The points lying on the lower boundary of the contours are most energy efficient for the given input and output constraints at given bulk-to-source voltage V_{BS} and represent the energy-delay curve of interest. Points on this curve can be

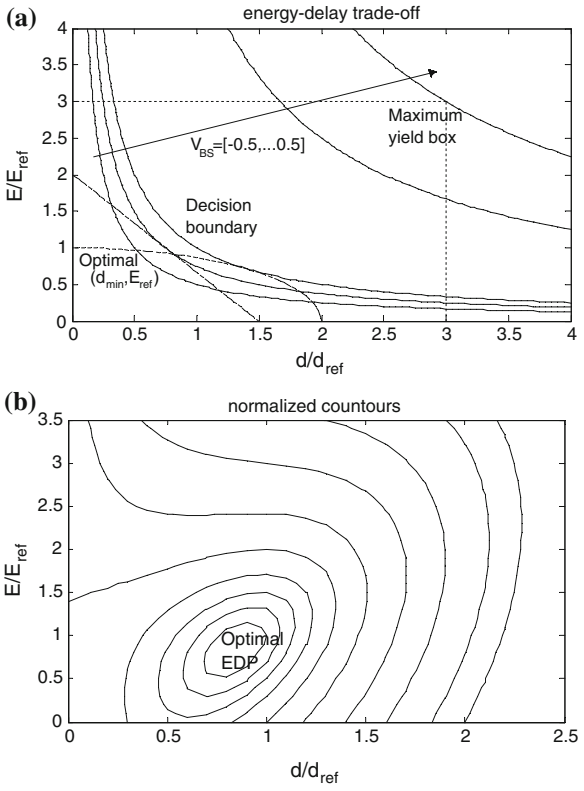

Fig. 2.4 **a** Optimal energy-delay tradeoff in a 64-bit adder obtained from energy minimization. Reference is the design sized for minimum delay under maximum allowed V_{DD} and reference V_T. **b** Normalized contours of energy showing optimal energy-delay product (EDP) point in $E/E_{ref} - d/d_{\text{ref}}$ plane

determined by sizing the circuit for minimal energy under the given input size and output load constraint for the desired delay target. This curve is often used for energy-delay tradeoff, where a design point is selected based on its cost of energy for a given change in delay. The reference design moves down on the *y*-axis to the optimal design point on the energy-efficient curve. With optimization, satisfying yield constrain, we can achieve energy savings of up to 55 % without any delay penalty. Alternatively, we can maintain the energy and achieve the speedup of about 25 %. Typically, only a subset of tuning variables d (e.g. gate size *W*, supply voltage V_{DD}, bulk-to-source voltage V_{BS}, etc.) is selected for optimization.

With a proper choice of the two variables, the designer can obtain nearly the minimal energy for a given delay. In our case, for delays close to d_{ref}, these variables are sizing and threshold voltage since there is the largest gap between the sizing and threshold voltage around the nominal delay point. The data in Fig. 2.4a shows that circuit optimization is really effective only in the region of about 30 % around the reference delay, d_{ref}. Outside this region, optimization becomes costly either in terms of delay or energy. Figure 2.4a also shows the decision boundary of the leakage energy corresponding to the minimal achievable energy-delay curve. The leakage curve is primarily affected by the large circuit size variation with respect to delay change. The increased leakage associated with a longer clock cycle is substantially less than the leakage reduction obtained from smaller transistor sizes. Therefore, leakage energy behaves as similarly as the active energy. Even when leakage energy becomes comparable to the active energy in future technologies or due to low switching activity of circuits, the characteristics of the minimal achievable energy-delay curve will remain unchanged and no algorithmic change for the optimization is needed. The obtained statistics of the total energy consumption for the benchmark circuit is compared with Monte Carlo based simulations. The results show that the estimates obtained using the proposed approach for the values of the mean delay and leakage energy are very accurate with an average error of 1.2 and 1.8 %, respectively. The standard deviations show an average error of 3.6 and 7.7 % for energy and delay, respectively.

Energy optimization for fixed input size and fixed output load: Energy optimization for a fixed input size and output load constraint is the most common design scenario. The plot in Fig. 2.4b illustrates the position of the optimal energy-delay product for 64-b static Kogge-Stone adder under maximum yield reference design point for the adder relative to the optimal energy-delay tradeoff curve obtained by jointly optimizing gate size, supply and threshold voltages. Through optimization, the input vectors are divided into a number of sub-sets. The optimization problem is solved incrementally, covering all the sub-sets of classes constructing the optimal separating hyperplane for the full data set. Note that during this process the value of the functional vector of parameters is monotonically increasing, since more and more training vectors are considered in the optimization leading to efficient separation between the two classes. In symmetrical circuit structures, the optimization space is limited and therefore the additional energy saving contributed by optimization is much smaller, especially with the higher timing yield. For decreased timing yield, higher energy saving can be

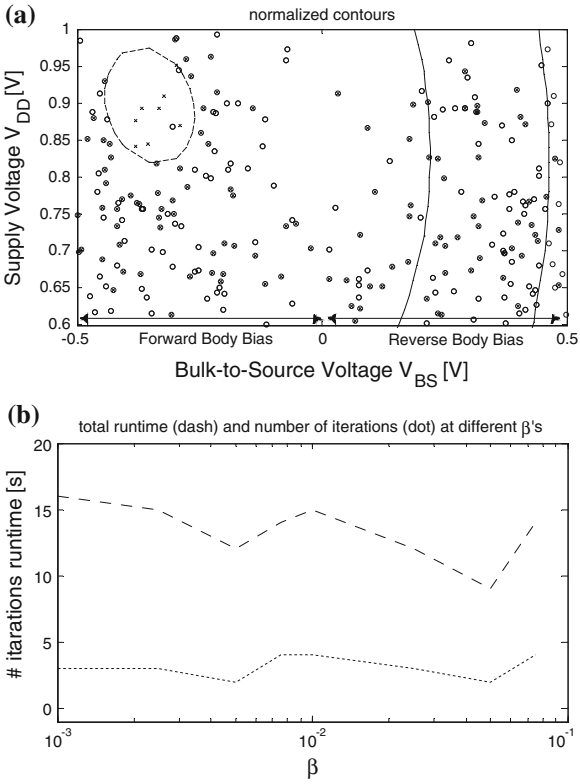

Fig. 2.5 **a** Normalized contours of energy in the $V_{DD} - V_{BS}$ plane of 64-b static Kogge-Stone adder. **b** Total runtime and number of iterations of 64-b static Kogge-Stone adder at different bound β

achieved as a consequence of a larger optimization space. Normalized contours in the $V_{DD} - V_{BS}$ plane are plotted in Fig. 2.5a. Monte Carlo simulations have been done to investigate an optimal operating region within which a circuit could function optimally and to verify its yield maximality. The total run-time of the statistical method (Fig. 2.5b) is only dozens of seconds, and the number of iterations required to reach the stopping criterion never exceeds 5 throughout the entire simulated β range (from 10^{-3} to 10^{-1}). Obtained optimum values for V_{DD} [V] are 0.855, 0.859, 0.862 and 0.877 and for V_{BS} [V] are -0.422, -0.408, -0.376 and -0.418 for Gaussian, non symmetric, highly kurtic and uniform distribution, respectively. Note in Fig. 2.5a, that bulk-to-source voltage (V_{BS}) modulates V_T, approach commonly used in practice. Any pair of V_{DD} and V_T in the feasible region satisfies the yield constraints for given E_{total}. In case when leakage energy dominates the total energy (e.g. low activity, high temperature), V_{BS} is increased to reduce the leakage. Resulting loss of performance is corrected by increasing V_{DD}. Similarly, when dynamic energy is dominant (e.g. high activity, low temperature), the total energy can be reduced by reducing V_{DD} and correcting the loss of performance by reducing V_{BS}. Note that the contours are normalized by dividing the minimum energy by the calculated energy for any pair of V_{DD} and V_{BS}, which

satisfy the yield constraints. To set tight constraints, the maximum allowed frequency can be lowered or the acceptable ratio of leakage to total power can be reduced. However, in an application for which activity of the circuit is high, the increase in the size of the transistors reduces the yield as a consequence of the increased transistors' parasitic capacitance. As yield increases when tolerance decreases, agreeable tradeoff needs to exist between increase in yield and the cost of design and manufacturing. Consequently, continuous observation of process variation and thermal monitoring becomes a necessity [104].

2.6 Conclusions

Statistical simulation is one of the foremost steps in the evaluation of successful high-performance IC designs due to process variations, which strongly affect devices behavior in today's deep submicron technologies. In this chapter, rather than estimating statistical behavior of the circuit by a population of realizations, we describe integrated circuits as a set of stochastic differential equations and introduce Gaussian closure approximations to obtain a closed form of moment equations. The static manufacturing variability and dynamic statistical fluctuation are treated separately. Process variations are modeled as a wide-sense stationary process and the solution of MNA for such a process is found. Similarly, we present a novel method to extend voltage-based gate models for statistical timing analysis. We constructed gate models based on statistical simplified transistor models for higher accuracy. Correlations among input signals and between input signal and delay are preserved during simulation by using same model format for the voltage and all elements in gate models. Furthermore, the multiple input simultaneous switching problem is addressed by considering all input signals together for output information. Since the proposed timing analysis is based on the transistor-level gate models, it is able to handle both combinational and sequential circuits. The experiments demonstrated the good combination of accuracy and efficiency of the proposed method for both deterministic and statistical timing analysis. Additionally, we present an efficient methodology for interconnect model reduction based on adjusted dominant subspaces projection. By adopting the parameter dimension reduction techniques, interconnect model extraction can be performed in the reduced parameter space, thus provide significant reductions on the required simulation samples for constructing accurate models. Extensive experiments are conducted on a large set of random test cases, showing very accurate results. Furthermore, we presented energy and yield constrained optimization as an active design strategy. We create a sequence of minimizations of the feasible region with iteratively-generated low-dimensional subspaces. As the resulting sub-problems are small, global optimization in both convex and non-convex cases is possible. The method can be used with any variability model, and is not restricted to any particular performance constraint. The effectiveness of the proposed approach is evaluated on a 64-b static Kogge-Stone adder implemented in UMC 1P8 M 65 nm

technology. As the experimental results indicate, the suggested numerical methods provide accurate and efficient solutions of energy optimization problem offering of up to 55 % energy savings.

References

1. K. Bowman, J. Meindl, Impact of within-die parameter fluctuations on the future maximum clock frequency distribution. *Proceedings of IEEE Custom Integrated Circuits Conference*, pp. 229–232 (2001)
2. T. Mizuno, J. Okamura, A. Toriumi, Experimental study of threshold voltage fluctuation due to statistical variation of channel dopant number in MOSFET's. IEEE Trans. Electron Devices **41**, 2216–2221 (1994)
3. A. Asenov, S. Kaya, J.H. Davies, Intrinsic threshold voltage fluctuations in MOSFETs due to local oxide thickness variations. IEEE Trans. Electron Devices **49**(1), 112–119 (2002)
4. J.A. Croon, G. Storms, S. Winkelmeier, I. Pollentier, Line-edge roughness: characterization, modeling, and impact on device behavior. *Proceedings of IEEE International Electronic Devices Meeting*, pp. 307–310 (2002)
5. A. Asenov, G. Slavcheva, A.R. Brown, J. Davies, S. Saini, Increase in the random dopant induced threshold fluctuations and lowering in sub-100 nm MOSFETs due to quantum effects: a 3-D density-gradient simulation study. IEEE Trans. Electron Devices **48**(4), 722–729 (2001)
6. J. Kwong, A. Chandrakasan, Variation driven device sizing for minimum energy subthreshold circuits. *IEEE International Symposium on Low-Power Electronic Design*, pp. 8–13 (2006)
7. M. Horowitz, E. Alon, D. Patil, S. Naffziger, R. Kumar, K. Bernstein, Scaling, power, and the future of CMOS. *IEEE International Electronic Devices Meeting*, pp. 7–15 (2005)
8. D. Markovic et al., Ultralow-power design in near-threshold region. Proc. IEEE **98**(2), 237–252 (2010)
9. K. Itoh, Adaptive circuits for the 0.5-V nanoscale CMOS era. *Digest of Techical Papers IEEE International Solid-State Circuits Conference,* pp. 14–20. (2009)
10. M. Grigoriu, On the spectral representation method in simulation. Probab. Eng. Mech. **8**, 75–90 (1993)
11. M. Loève, *Probability Theory* (D. Van Nostrand Company Inc., Princeton, 1960)
12. R. Ghanem, P.D. Spanos, *Stochastic Finite Element: A Spectral Approach* (Springer, New York, 1991)
13. P. Friedberg, Y. Cao, J. Cain, R. Wang, J. Rabaey, C. Spanos, Modeling within-die spatial correlation effects for process-design co-optimization. *IEEE International Symposium on Quality of Electronic Design*, pp. 516–521 (2005)
14. J. Xiong, V. Zolotov, L. He, Robust extraction of spatial correlation. *Proceedings of IEEE International Symposium on Physical Design*, pp. 2–9 (2006)
15. M. Pelgrom, A. Duinmaijer, A. Welbers, Matching properties of MOS transistors. IEEE J. Solid-State Circuits **24**(5), 1433–1439 (1989)
16. C. Michael, M. Ismail, *Statistical Modeling for Computer-Aided Design of MOS VLSI Circuits* (Kluwer, Boston, 1993)
17. H. Zhang, Y. Zhao, A. Doboli, ALAMO: an improved σ-space based methodology for modeling process parameter variations in analog circuits. *Proceedings of IEEE Design, Automation and Test in Europe Conference*, pp. 156–161 (2006)
18. R. López-Ahumada, R. Rodríguez-Macías, FASTEST: a tool for a complete and efficient statistical evaluation of analog circuits, dc analysis. Analog Integr. Circ. Sig. Process. **29**(3), 201–212 (2001)(Kluwer Academic Publishers)

19. G. Biagetti, S. Orcioni, C. Turchetti, P. Crippa, M. Alessandrini, SiSMA-a statistical simulator for mismatch analysis of MOS ICs. *Proceedings of IEEE/ACM International Conference on Computer Aided Design*, pp. 490–496 (2002)
20. B. De Smedt, G. Gielen, WATSON: design space boundary exploration and model generation for analogue and RF IC design. IEEE Trans. CAD Integr. Circuits Syst. **22**(2), 213–224 (2003)
21. B. Linares-Barranco, T. Serrano-Gotarredona, On an efficient CAD implementation of the distance term in Pelgrom's mismatch model. IEEE Trans. CAD Integr. Circuits Syst. **26**(8), 1534–1538 (2007)
22. J. Kim, J. Ren, M.A. Horowitz, Stochastic steady-state and ac analyses of mixed-signal systems. *Proceedings of IEEE Design Automation Conference*, pp. 376–381 (2009)
23. A. Zjajo, J. Pineda de Gyvez, Analog automatic test pattern generation for quasi-static structural test. IEEE Trans. VLSI Syst. **17**(10), 1383–1391 (2009)
24. N. Mi, J. Fan, S.X.-D. Tan, Y. Cai, X. Hong, Statistical analysis of on-chip power delivery networks considering lognormal leakage current variations with spatial correlation. IEEE Trans. Circuits Syst. I Regul. Pap. **55**(7), 2064–2075 (2008)
25. E. Felt, S. Zanella, C. Guardiani, A. Sangiovanni-Vincentelli, Hierarchical statistical characterization of mixed-signal circuits using behavioral modeling. *Proceedings of IEEE Inernational Conference on Computer Aided Design,* pp. 374–380 (1996)
26. J. Vlach, K. Singhal, *Computer Methods for Circuit Analysis and Design* (Van Nostrand Reinhold, New York, 1983)
27. L.O. Chua, C.A. Desoer, E.S. Kuh, *Linear and Nonlinear Circuits* (Mc Graw-Hill, New York, 1987)
28. L. Arnold, *Stochastic Differential Equations: Theory and Application* (Wiley, New York, 1974)
29. S. Bhardwaj, S. Vrudhula, A. Goel, A unified approach for full chip statistical timing and leakage analysis of nanoscale circuits considering intradie process variations. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. **27**(10), 1812–1825 (2008)
30. J.F. Croix, D.F. Wong, A fast and accurate technique to optimize characterization tables for logic sythesis. *Proceedings of IEEE Design Automation Conference*, pp. 337–340 (1997)
31. A. Goel, S. Vrudhula, Statistical waveform and current source based standard cell models for accurate timing analysis. *Proceedings of IEEE Design Automation Conference*, pp. 227–230 (2008)
32. H. Fatemi, S. Nazarian, M. Pedram, Statistical logic cell delay analysis using a current-based model. *Proceedings of IEEE Design Automation Conference*, pp. 253–256 (2006)
33. B. Liu, A.B. Kahng, Statistical gate level simulation via voltage controlled current source models. *Proceedings of IEEE International Workshop on Behavioral Modeling and Simulation*, p. 23–27 (2006)
34. B. Liu, Gate level statistical simulation based on parameterized models for process and signal variations. *Proceedings of IEEE International Symposium on Quality Electronic Design*, pp. 257–262 (2007)
35. J.F. Croix, D.F. Wong, Blade and Razor: cell and interconnet delay analysis using current-based models. *Proceedings of IEEE Design Automation Conference*, pp. 386–389 (2003)
36. C. Amin, C. Kashyap, N. Menezes, K. Killpack, E. Chiprout, A multi-port current source model for multiple-input switching effects in CMOS library cells. *Proceedings of IEEE Design Automation Conference*, pp. 247–252 (2006)
37. C. Kashyap, C. Amin, N. Menezes, E. Chiprout, A nonlinear cell macromodel for digital applications. *Proceedings of IEEE International Conference on Computer Aided Design*, pp. 678–685 (2007)
38. N. Menezes, C. Kashyap, C. Amin, A true electrical cell model for timing, noise, and power grid verification. *Proceedings of IEEE Design Automation Conference*, pp. 462–467 (2008)
39. B. Amelifard, S. Hatami, H. Fatemi, M. Pedram, A current source model for CMOS logic cells considering multiple input switching and stack effect. *Proceedings of IEEE Design, Automation and Test in Europe Conference*, pp. 568–574 (2008)

40. A. Devgan, Accurate device modeling techniques for efficient timing simulation of integrated circuits. *Proceedings of IEEE International Conference on Computer Design*, pp. 138–143 (1995)
41. F. Dartu, Gate and transistor level waveform calculation for timing analysis. Ph.D. Dissertation, Carnegie Mellon University, 1997
42. P. Kulshreshtha, R. Palermo, M. Mortazavi, C. Bamji, H. Yalcin, Transistor-level timing analysis using embedded simulation. *Proceedings of IEEE International Conference on Computer Aided Design*, pp. 344–349 (2000)
43. P.F. Tehrani, S.W. Chyou, U. Ekambaram, Deep sub-micron static timing analysis in presence of crosstalk. *Proceedings of IEEE International Symposium on Quality Electronic Design*, pp. 505–512 (2000)
44. E. Acar, Linear-centric simulation approach for timing analysis. Ph.D. dissertation, Carnegie Mellon University, 2001
45. E. Acar, F. Dartu, L. Pileggi, TETA: transistor-level waveform evaluation for timing analysis. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. **21**(5), 605–616 (2002)
46. L. McMurchie, C. Sechen, WTA-waveform-based timing analysis for deep-micro circuits. *Proceedings of IEEE International Conference on Computer Aided Design*, pp. 625–631 (2002)
47. Z. Wang, J. Zhu, Transistor-level static timing analysis by piecewise quadratic waveform matching. *Proceedings of IEEE Design, Automation and Test in Europe Conference*, pp. 312–317 (2003)
48. S. Raja, Varadi, M. Becer, J. Geada, Transistor level gate modeling for accurate and fast timing, noise, and power analysis. *Proceedings of IEEE Design Automation Conference*, pp. 456–461 (2008)
49. Q. Tang, A. Zjajo, M. Berkelaar, N. van der Meijs, Transistor level waveform evaluation for timing analysis. in *Proceedings of European Workshop on CMOS Variability*, pp. 1–6 (2010)
50. J.F. Epperson, *An Introduction to Numerical Methods and Analysis* (John Wiley & Sons, Inc, New York, 2002)
51. T. Shima, H. Yamada, R.L.M. Dang, Table look-up mosfet modeling system using a 2-d device simulator and monotonic piecewise cubic interpolation. IEEE Trans. Comput. Aided Des. **2**(2), 121–126 (1983)
52. P.E. Allen, K.S. Yoon, A table look-up model for analog applications. *International Conference on Computer-Aided Design*, pp. 124–127 (1988)
53. *Pathmill: Transistor-level static timing analysis*, [online], available at: http://www.synopsys.com/products/analysis/pathmillds.pdf
54. Q. Tang, A. Zjajo, M. Berkelaar, N. van der Meijs, A simplified transistor model for cmos timing analysis. *Proceedings of Workshop on circuits, systems and signal processing*, pp. 289–294 (2009)
55. M. Chen, W. Zhao, F. Liu, Y. Cao, Fast statistical circuit analysis with finite-point based transistor model. *Proceedings of IEEE Design, Automation and Test in Europe Conference*, pp. 1–6 (2007)
56. A. Hyvarinen, E. Oja, Independent component analysis: algorithms and applications. Neural Networks J. **13**(4/5), 411–430 (2000)
57. R. Manduchi, J. Portilla, Independent component analysis of textures. Proc. IEEE Int. Conf. Comput. Vis. **2**, 1054–1060 (1999)
58. Z. Feng, P. Li, Y. Zhan, Fast second-order statistical static timing analysis using parameter dimension reduction. *Proceedings of IEEE Design Automation Conference*, pp. 244–249 (2007)
59. C. Visweswariah et al., First-order incremental block-based statistical timing analysis. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. **25**(10), 2170–2180 (2006)
60. T.T. Soong, *Random Diffrential Equations in Science and Engineering* (Academic Press, New York, 1973)

61. Q. Tang, A. Zjajo, M. Berkelaar, N. P. van der Meijs, RDE-based transistor-level gate simulation for statistical static timing analysis. *Proceedings of IEEE Design Automation Conference,* pp. 787–792 (2010)
62. Q. Tang, A. Zjajo, M. Berkelaar, N.P. van der Meijs, Statistical delay calculation with multiple input simultaneous switching. *Proceedings of IEEE International Conference on IC Design and Technology*, pp. 1–4 (2011)
63. L.T. Pillage, R.A. Rohrer, Asymptotic waveform evaluation for timing analysis. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. **4**, 352–366 (1990)
64. P. Feldmann, R.W. Freund, Efficient linear circuit analysis by Pade approximation via the Lanczos process. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. **14**, 639–649 (1995)
65. A. Odabasioglu, M. Celik, L. Pileggi, PRIMA: Passive reduced-order interconnect macromodeling algorithm. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, pp. 645–654 (1998)
66. P. Elias, N. van der Meijs, Including higher-order moments of RC interconnections in layout-to-circuit extraction. *Proceedings of IEEE Design, Automation and Test in Europe Conference*, pp. 362–366 (1996)
67. B.C. Moore, Principal component analysis in linear systems: controllability, observability, and model reduction. IEEE Trans. Autom. Control **26**, 17–31 (1981)
68. J. Li, J. White, Efficient model reduction of interconnect via approximate system Grammians. *Proceedings of IEEE International Conference on Computer Aided Design*, pp. 380–384 (1999)
69. J.R. Phillips, L. Daniel, L.M. Silveira, Guaranteed passive balancing transformations for model order reduction. *Proceedings of IEEE Design Automation Conference*, pp. 52–57 (2002)
70. J.R. Phillips, L.M. Silveira, Poor man's TBR: a simple model reduction scheme. *Proceedings of IEEE Design, Automation and Test in Europe Conference*, pp. 938–943 (2004)
71. W.F. Arnold, A.J. Laub, Generalized eigenproblem algorithms and software for algebraic Riccati equation. Proc. IEEE **72**, 1746–1754 (1984)
72. T. Penzl, A cyclic low-rank Smith method for large sparse Lyapunov equations. SIAM J. Sci. Comput. **21**, 1401–1418 (2000)
73. M.G. Safonov, R.Y. Chiang, A Schur method for balanced-truncation model reduction. IEEE Trans. Autom. Control **34**, 729–733 (1989)
74. K.V. Fernando, H. Nicholson, Singular perturbational model reduction of balanced systems. IEEE Trans. Autom. Control **27**, 466–468 (1982)
75. D. Enns, Model reduction with balanced realizations: an error bound and a frequency weighted generalization. *Proceedings of IEEE Conference on Decision and Control*, pp. 127–132 (1984)
76. M.S. Tombs, I. Postlethwaite, Truncated balanced realization of stable, non-minimal state-space systems. Int. J. Control **46**, 1319–1330 (1987)
77. G. Golub, C. van Loan, *Matrix Computations* (Johns Hopkins University Press, Baltimore MD, 1996)
78. J. Singh, V. Nookala, Z. Luo, S. Sapatnekar, Robust gate sizing by geometric programming. *Proceedings of IEEE Design Automation Conference*, pp. 315–320 (2005)
79. D. Nguyen et al., Minimization of dynamic and static power through joint assignment of threshold voltages and sizing optimization. *Proceedings of IEEE International Symposium on Low Power Electronic Design*, pp. 158–163 (2003)
80. R. Brodersen et al., Methods for true power minimization. *Proceedings of IEEE International Conference on Computer-Aided Design*, pp. 35–42 (2002)
81. K. Nose, T. Sakurai, Optimization of VDD and VTH for low power and high-speed applications. *Proceedings of IEEE Design Automation Conference*, pp. 469–474 (2000)
82. A. Bhavnagarwala, B. Austin, K. Bowman, J.D. Meindl, A minimum total power methodology for projecting limits on CMOS GSI. IEEE Trans. VLSI Syst. **8**(6), 235–251 (2000)

83. M. Mani, A. Devgan, M. Orshansky, An efficient algorithm for statistical minimization of total power under timing yield constraints. *Proceedings of IEEE Design Automation Conference*, pp. 309–314 (2005)
84. A. Srivastava, K. Chopra, S. Shah, D. Sylvester, D. Blaauw, A novel approach to perform gate-level yield analysis and optimization considering correlated variations in power and performance. IEEE Trans. Comput. Aided Des. **27**(2), 272–285 (2008)
85. C. Gu, J. Roychowdhury, An efficient, fully nonlinear, variability-aware non-Monte-Carlo yield estimation procedure with applications to SRAM cells and ring oscillators. *Proceedings of IEEE Asia-South Pacific Design Automation Conference*, pp. 754–761 (2008)
86. M. Meijer, J. Pineda de Gyvez, Body bias driven design synthesis for optimum performance per area. *Proceedings of IEEE International Symposium on Quality Electronic Design*, pp. 472–477 (2010)
87. A. Zjajo, Q. Tang, M. Berkelaar, J. Pineda de Gyvez, A. Di Bucchianico, N. van der Meijs, Stochastic analysis of deep-submicrometer CMOS process for reliable circuits designs. IEEE Trans. Circuits Syst. I Regul. Pap. **58**(1), 164–175 (2011)
88. Y. Freund, R.E. Schapire, Large margin classification using the perceptron algorithm. Mach. Learn. **37**, 277–296 (1999)
89. I. Tsochantaridis, T. Hofmann, T. Joachims, Y. Altun, Support vector machine learning for interdependent and structured output spaces. *Proceedings of of International Conference on Machine Learning*, pp. 1–8 (2004)
90. J.C. Platt, Fast training of support vector machines using sequential minimal optimization, in *Advances in Kernel Methods: Support Vector Learning,* ed. by B. Scholkopf, C.J.C. Burges, A.J. Smola (MIT Press, Cambridge, 1998) pp. 195–208
91. B. Taskar, Learning structured prediction models: a large margin approach, PhD thesis, Stanford University, 2004
92. V. Franc, V. Hlavac, Multi-class support vector machine. Proc. IEEE Int. Conf. Pattern Recognit. **2**, 236–239 (2002)
93. MatLab, http://www.mathworks.com/
94. A. Zjajo, M. Song, Digitally programmable continuous-time biquad filter in 65-nm CMOS. *Proceedings of IEEE International Symposium on Radio-Frequency Integration Technology*, pp. 339–342 (2009)
95. Nangate 45 nm open cell library (2009), http://www.nangate.com/ index.php?option = comcontent&task = view&id = 137&Itemid = 137
96. X. Lu, W.P. Shi., Layout and parasitic information for ISCAS circuits (2004), http://dropzone.tamu.edu/xiang/iscas.html
97. X. Zheng, Implementing and evaluating a simplified transistor model for timing analysis of integrated circuits, Master's thesis, Delft University of Technology, 2012
98. J. Rodriguez, Q. Tang, A. Zjajo, M. Berkelaar, N. van der Meijs, Direct statistical simulation of timing properties in sequential circuits. *Proceedings of International Workshop on Power and Timing Modeling, Optimization ans Simulation*, pp. 131–141 (2012)
99. MiBench, http://www.eecs.umich.edu/mibench/
100. SimpleScalar, http://www.simplescalar.com/
101. HSPICE Simulation and Analysis User Guide, Version W-2005.03, Synopsys, Mountain View, CA, 2005
102. P.M. Kogge, H.S. Stone, A parallel algorithm for the efficient solution of general class of recurrence equations. IEEE Trans. Comput. **C-22**(8), 786–793 (1973)
103. K. Bernstein et al., High-performance CMOS variability in the 65 nm regime and beyond. IBM J. Res. Dev. **50**(4/5), 433–449 (2006)
104. A. Zjajo, M.J. Barragan, J. Pineda de Gyvez, Low-power die-level process variation and temperature monitors for yield analysis and optimization in deep-submicron CMOS. IEEE Trans. Instrum. Meas. **61**(8), 2212–2221 (2012)

http://www.springer.com/978-94-007-7780-4

Stochastic Process Variation in Deep-Submicron CMOS
Circuits and Algorithms

Zjajo, A.

2014, XIX, 192 p. 46 illus., Hardcover
ISBN: 978-94-007-7780-4