Overcoming Variations in Nanometer-Scale Technologies

Sachin S. Sapatnekar, *Fellow, IEEE*

***Abstract*— Nanometer-scale circuits are fundamentally different from those built in their predecessor technologies in that they are subject to a wide range of new effects that induce on-chip variations. These include effects associated with printing finer geometry features, increased atomic-scale effects, and increased on-chip power densities, and are manifested as variations in process and enviromental parameters and as circuit aging effects. The impact of such variations on key circuit performance metrics is quite significant, resulting in parametric variations in the timing and power, and potentially catastrophic failure due to reliability and aging effects. Such problems have led to a revolution in the way that chips are designed in the presence of such uncertainties, both in terms of performance analysis and optimization. This paper presents an overview of the root causes of these variations and approaches for overcoming their effects.**

I. Introduction

The effects of variability in nanometer-scale integrated circuits cause significant deviations from the prescribed specifications for a chip. The magnitude of these deviations, together with tight performance specifications, imply that variability is an increasingly vexing problem as technologies continue to scale.

The sources of these variations can be categorized into several classes, depending on their origin:

- *Process variations* are one-time variations that occur when a circuit is manufactured, and cause process parameters to drift from their designed values.
- *Environmental variations* are run-time variations that reflect the effects of altered operating conditions during the operation of a circuit. Such variations may be attributed to factors such as supply voltage changes, thermal effects, and radiation-induced soft errors.
- *Aging variations* reflect the fact that the behavior of a circuit degrades as it ages, due to the prolonged application of stress. Such degradations may result in parametric degradations or catastrophic failures.

These variations can impact key circuit performance characteristics: for digital circuits, the affected parameters include the delay, power, and lifetime of the circuit, while for analog circuits, the performance parameters to be monitored are specific to the type of circuit.

It is increasingly obvious that designing circuits at the nominal point, or using simple corner-based approaches, is no longer viable. The field of robust design, which was largely confined to analog circuits in the past, has now become an integral part of the design process for digital circuits as well. Even for analog parts, the underlying causes of variation have been altered to the extent that a fresh look must be taken at design for variability. Such factors are being exacerbated by the emergence of new technologies such as three-dimensional integrated circuits (3DICs), where variations due to power and thermal issues can be very significant.

The nominal supply voltage is a significant factor in determining the extent of circuit performance variations. Superthreshold circuits, which constitute the mainstream of today's designs, use a supply voltage that is significantly larger than the transistor threshold voltage. These circuits see significant shifts in the leakage power, where exponential factors come into play. The variations in delay, of a few tens of a percent, seem superficially more moderate, but constitute extremely large variations given the tight specifications that the circuits are designed to satisfy, and the expense associated with allocating on-chip resources to bring the circuits back to specifications.

In contrast, subthreshold or near-threshold circuits set the supply voltage to be, respectively, below or just above the transistor threshold voltage to achive significant gains in power and/or energy efficiency. In this regime, delays are also ruled by expressions that involve exponentials, and therefore the magnitude of these shifts can be very significant. While a significant amount of research has been carried out into the analysis and optimization of superthreshold circuits, as outlined in this paper, the treatment of subthreshold and near-threshold circuits largely remains an open problem.

II. Process variations

A. Sources of process variation

Examples of variations during the manufacturing process include shifts in the values of parameters such as the effective channel length (L_{eff}), the oxide thickness (t_{ox}), the dopant concentration (N_a), the transistor width (w), the interlayer dielectric (ILD) thickness (t_{ILD}), and the interconnect height and width (h_{int} and w_{int}, respectively). Examples of such variations are illustrated in Figure 1 [1]–[3].

Process variations can be classified into the following categories, depending on their physical range on a die or wafer:

- *Die-to-die (D2D) variations* correspond to changes from one die to another (Figure 2(a)).
- *Within-die (WID) variations* correspond to variability within a single die (Figure 2(b)).

S. S. Sapatnekar is with the Department of Electrical and Computer Engineering, Univeristy of Minnesota, Minneapolis, MN 55455, USA. This research was supported in part by the SRC under contract 2007-TJ-1572 and by the NSF under awards CCF-0634802 and CCF-1017778.

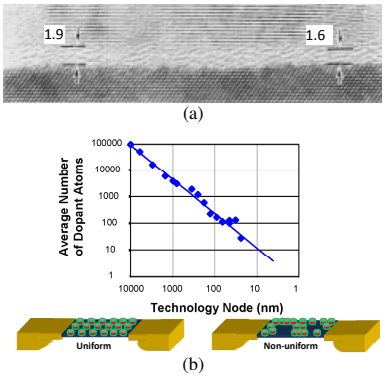

Fig. 1. Examples of process variations in (a) gate oxide thickness and (b) the number and distribution of dopant atoms in a transistor.

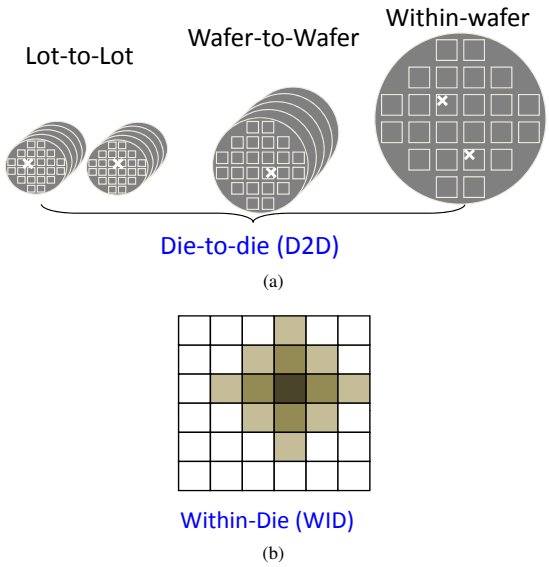

Fig. 2. A taxonomy of variations, illustrating (a) various types of die-to-die variations and (b) within-die variations.

D2D variations affect all the devices on same chip in the same way, e.g., causing the transistor gate lengths of devices on the same chip to be all larger or all smaller than the nominal, while WID variations may affect different devices differently on the same chip, e.g., causing some devices have smaller transistor gate lengths and others larger transistor gate lengths than the nominal. D2D variations have been a longstanding design issue, and for several decades, designers have striven to make their circuits robust under the unpredictability of such variations. This has typically been achieved by simulating the design at not just one design point, but at a small number of "corners." These corners are chosen to encapsulate the behavior of the circuit under worst-case variations, and have served designers well in the past. In nanometer technologies, WID variations have become significant and can no longer be ignored. Corner-based methods are adequate in the case where all variations are D2D, and no WID variations are seen.

The sources of on-chip variations may be used to create another taxonomy:

- *Systematic variations* show predictable variational trends across a chip, and are caused by known physical phenomena during manufacturing.
- *Random variations* depict random behavior that can be characterized in terms of a distribution.

Systematic variations can arise due to several factors, e.g., due to (i) spatial WID gate length variability, also known as across-chip linewidth variation (ACLV), which are manifested as systematic changes in the value of L_{eff} across a reticle due to changes in the stepper-induced illumination, imaging nonuniformity due to lens aberrations, etc., and (ii) ILD variations due to the effects of chemical-mechanical polishing (CMP) on metal density patterns: regions that have uniform metal densities tend to have more uniform ILD thicknesses than regions that have nonuniformities. A more detailed discussion of systematic variations, their effects, and their optimization is beyond the scope of this paper. It is important to note that random variations include those whose origins can be truly said to be random (e.g., random dopant fluctuations) as well as those that are not truly random, but that are difficult to model as systematic variations.

Random variations are associated with a probability distribution that may either be explicit, in terms of a large number of samples provided from fabrication line measurements, or implicit, in terms of a known probability density function (such as a Gaussian or a lognormal distribution) that has been fitted to the measurements. Random variations in some process or environmental parameters (such as those in the temperature, supply voltage, or L_{eff}) can often show a degree of local spatial correlation, whereby variations in one transistor in a chip are remarkably similar in nature to those in spatially neighboring transistors, but may differ significantly from those that are far away. Other process parameters (such as t_{ox} and N_a) do not show much spatial correlation at all, so that for all practical purposes, variations in neighboring transistors are uncorrelated.

Some, but not all, randomly varying parameters show spatial correlations, where the probability density functions show correlations related on the spatial location of objects. A classical model for spatial correlation, which predicts the decay with distance, was proposed by Pelgrom [4]. For the purposes of statistical analyses, more approximate models that capture the spirit of these distance-based variations are adequate. For instance, commonly-used models [5]–[7] tessellate the die into n grid region, with the values of a parameter within a grid being perfectly correlated, while the correlations between parameters in different grids depend on the distance between the grids. Other approaches use a continuous correlation model based on the Kosambi-Karhunen-Loève expansion [8]. The characterization of spatial correlations has been studied in works such as [9], [10].

Correlations affect the results of analysis of timing and power. For example, spatially uncorrelated variations tend to see large degrees of cancellation of randomness. Spatially correlated variations do not permit this cancellation, since in a region of the chip, most transistor parameters trend in the same directions, leaving fewer possibilities for such averaging. Therefore, correlations tend to exaggerate variations, and

performance simulations that model correlation tend to show wider variances than those that ignore them.

In addition to spatial correlations, circuits may show structural correlations that affect their timing behavior. For example, if two paths in a circuit include a common set of gates, the path delay function, which is the sum of the gate delays, must clearly show some correlation due to the contribution of the random delays associated with the common gates.

B. Analysis and optimization of process variations

1) Timing analysis: In the presence of manufacturing variations, the underlying economic model dictates the design objective: for microprocessors, where performance variations are typically dealt with by binning, and slower or faster processors are sold for lower or higher prices, respectively; the objective is to maximize profit, which can be translated into a minimum target yield for each bin. Under the ASIC model, binning is less prevalent and design constraints can be tight: a design either meets them or does not. Such a scenario is less forgiving, as compared to the binning model, of performance shifts due to variations, and statistical design can be of even greater utility.

Statistical static timing analysis (SSTA) and statistical power analysis represent the generalization of traditional corner-based static timing analysis (STA) and power estimation techniques, respectively. These methods treat circuit performance metrics, such as delay and power, not as fixed numbers, but as probability density functions (PDFs), taking the statistical distribution of parametric variations into consideration while analyzing the circuit. The simplest way to achieve this, in terms of the complexity of implementation, may be through Monte Carlo analysis. Monte Carlo analysis generates samples of the variational parameters, either according to raw data or based on the underlying PDF, and simulates the performance of the circuit. The histogram of the performance over a sufficiently large number of sample serves as an approximation to its PDF. While such an analysis can handle arbitrarily complex variations, its major disadvantage is in its extremely large runtimes. Therefore, more efficient methods are called for, based on SSTA.

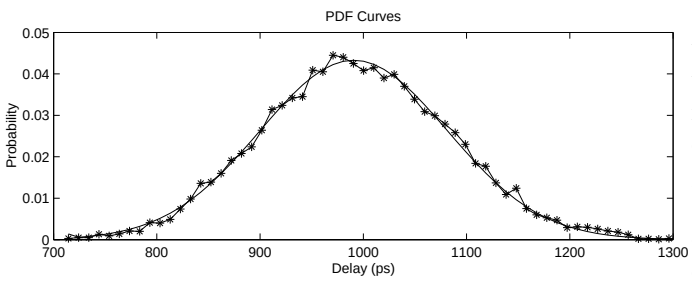

Fig. 3. The PDF for the clock period of circuit s38417. The curve marked by the solid line denotes the results of the SSTA engine, *MinnSSTA*, while the plot marked by the starred lines denotes the results of *MC*.

Figure 3 illustrates the results of SSTA on the benchmark circuit s38417, using Monte Carlo simulation (*MC*) as well as with an analytic method (*MinnSSTA* [5]). This example provides a proof of concept of the idea that analytical methods can be used to supplant Monte Carlo analysis and perform SSTA accurately and efficiently.

SSTA begins with a typical variational model of the delay of a gate in the form of a representation, $D = f(\mathbf{p})$, where $\mathbf{p}$ is the set of underlying process parameters. For small variations in the p_i variables, the delay function can be expressed in the form of a first-order or second-order Taylor series expansion. A second-order expansion has the form:

$$D = D_0 + \sum_i \left[\frac{\partial f}{\partial p_i}\right]_0 \Delta p_i + \sum_i \sum_j \left[\frac{\partial^2 f}{\partial p_i \partial p_j}\right]_0 \Delta p_i \Delta p_j, \quad (1)$$

where D_0 is the nominal value of D, calculated at the nominal values of parameters in $\mathbf{p}$, the first and second partial derivatives are is computed at the nominal values of p_i, and $\Delta p_i = p_i - E[p_i]$ (where $E[.]$ is the expectation operator) is a zero-mean random variable representing parameter variations about the nominal values. SSTA uses these gate delay models and propagates the delay PDFs to the circuit outputs.

Mainstream approaches to SSTA use block-based methods, which compute delays using a PERT-like topological traversal of a circuit, as against path-based methods that enumerate paths. Most methods are based on continuous probability density functions. SSTA techniques can further be classified into methods that use:

- Gaussian vs. non-Gaussian modeling: This classification corresponds to the PDF used to represent the underlying variations. If the underlying parameters $\Delta p_i \in \mathbf{p}$ in (1) are all random variables with a Gaussian distribution, then D is a linear combination of normally distributed random variables, and its PDF is Gaussian.
- Linear vs. nonlinear analysis: In the presence of variations, the Taylor series representation of the delay, about the nominal point, can be a truncated first order representation, as in (1). If the variations are small, this linear expansion is adequate; in case of larger variations, higher order nonlinear terms (typically quadratic) must be introduced.

The task of static timing analysis involves a topological traversal across a combinational circuit, processing each gate to determine its output arrival times after all information about its input arrival times is known [11]. STA operations can be distilled into two types: the "sum" and "max" operations. A gate is processed in STA when the arrival times of all inputs are known, at which time the candidate delay values at the output are computed using the sum operation, which adds the delay at each input with the input-to-output pin delay. Once these candidate delays have been found, the max operation is applied to determine the maximum arrival time at the output. In SSTA, the operations are identical to STA; the difference is that the pin-to-pin delays and the arrival times are PDFs instead of single numbers.

In classifying SSTA methods, we consider all four combinations of the above two bullets below:

The linear/Gaussian case: Under the Gaussian assumption, the dominant paradigm for SSTA uses principal component analysis (PCA) [12] to orthogonalize a set of correlated Gaussians in a transformed space. This step is typically performed as a preprocessing step that is computed once for a given technology. Efficient methods for computing the sum and max functions are outlined in [5], [13], [14], allowing for efficient SSTA for the linear/Gaussian case in the presence of correlated and uncorrelated variations.

The nonlinear/Gaussian case: For the nonlinear/Gaussian case a moment-based approach can be employed [15], [16]. The circuit delay function is modeled, using a response surface modeling approach, as a quadratic function of the process parameters. Correlated parameters are first orthogonalized using principal components analysis, and then a diagonalization approach is used to transform the quadratic function to remove cross-terms of the type $\Delta p_i \Delta p_j$. A key property of this diagonalization is that it preserves the orthogonality of the principal components. This work was subsequently extended [17] to develop a clever set of manipulations to compute the result of the max operator.

The linear/non-Gaussian case: For this case, the best solution to date is an approach [18], [19] that transforms Gaussian parameter PDFs using PCA, and orthogonalizes non-Gaussian parameter PDFs using a procedure known as independent component analysis (ICA) [20], provides an efficient solution. All parameter PDFs are represented in terms of their moments, which are used to obtain the moments of the orthogonalized PDFs in a preprocessing step. These are then propagated through the circuit to obtain the delay PDF for the circuit.

The nonlinear/non-Gaussian case: The nonlinear non-Gaussian case covers the most general case for performing statistical timing analysis. Several approaches [21]–[24] to this problem have been presented, but they all rely on computationally expensive techniques that are not scalable to a large number of correlated variables. Although quadratic models may be used and orthogonalized, similar to the nonlinear/Gaussian case in [15], [16], the ICA transform that applies them to orthogonalized non-Gaussians [18], [19] can only guarantee that the PDFs in the transformed space will be uncorrelated, but not that they will be independent. This limitation hinders the computation of higher-order moments for non-Gaussians. The quest for an efficient SSTA technique for this problem remains an open research problem.

An alternative class of approaches to timing analysis under variations uses statistically-based methods to extend the corner-based paradigm to a statistical design scenario [25], or attempt pessimistic worst-case modeling [26], [27]. Such methods aim to preserve the classic corner-based methods, but use intelligent statistical methods to determine the corners.

2) Power analysis: The power dissipation of a component is composed of three components: the dynamic power, the short-circuit power, and the leakage power. Of these, the first two are not especially sensitive to variations. Leakage power is related to several process parameters through exponential relationships, and therefore, a small parameter change can cause a large change in the leakage. Since leakage forms a large portion of the total power in nanometer-scale technologies, any variations can significantly impact the total power dissipation of a chip.

The major components of leakage in current CMOS technologies are due to sub-threshold leakage and gate tunneling leakage. The analysis of total leakage power of circuit is complicated by the state dependency of subthreshold and gate tunneling leakage, and the interactions between these two leakage mechanisms [28]. Other work [29], [30] presents an analytical framework that provides a closed form expression for the total chip leakage current as a function of process parameters for uncorrelated variations. This is used to estimate yield under power and performance constraints.

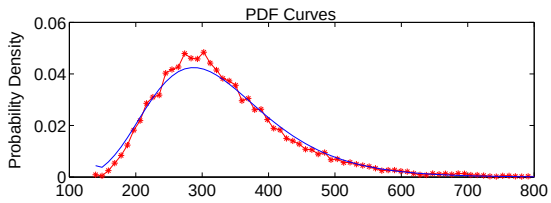

Fig. 4. The PDF for the power dissipation of circuit s38417. The curve marked by the solid line denotes the results of an analytical approach based on lognormal models, while the plot marked by the starred lines denotes the results of *MC*.

A key observation is that the subthreshold leakage can be written as an exponential function of L_{eff}. Under process variations, a linear approximation of this function may be used, as in SSTA. The first order Taylor series expansion of Gaussian parameter variations yields a Gaussian, and when these are exponentiated, the resulting distribution is lognormal [31]. Similarly, the gate leakage can be written as an exponential function of T_{ox}, and also yields a lognormal distribution for a gate. Under the assumption that all variations are independent, the sum of the leakage of all gates in a circuit approaches a normal distribution under the central limit theorem; when this sum is taken over a million or a billion transistors, the variance is negligible, and the leakage is characterized by a mean that can be calculated analytically [30]. The sum of lognormals can be approximated as a lognormal using Wilkinson's method [32]; the complexity of this method is linear in the number of terms to be added when the PDFs are uncorrelated, but quadratic in the presence of correlation. Other work [33] presents an approach for efficiently performing the addition using Wilkinson's method by reducing the effects of cross-terms. Another approach [34] uses the PCA orthogonalization of the original parameters to ensure that Wilkinson's method can work with uncorrelated PDFs. These two methods may be hybridized [35], and the resulting approach is shown to be better than either one individually. The results of applying this class of methods to compute the PDF of the power dissipation is illustrated in Figure 4: it is easily seen that the analytical approach provides an excellent approximation to the more exact Monte Carlo-based computation.

C. Optimization

Process variations can significantly degrade the yield of a circuit, and optimization techniques can be used to improve the timing yield. An obvious way to increase the timing yield of the circuit is to pad the specifications to make the circuit robust to variations, i.e., to choose a delay specification of the circuit that is tighter than the required delay. This new specification must be appropriately selected to avoid large area or power overheads due to excessively conservative padding.

Several techniques for timing yield optimization using gate sizing have been published in the literature. This step is performed close to the point where the layout of the circuit is known, and therefore, the design uncertainty due to unknown parasitics at this stage is relatively low; in contrast, in early parts of the design flow, the design uncertainty may overshadow any benefits that may be predicted by optimization. Early work [36], proposes formulation of statistical objective and timing constraints, and solves the resulting nonlinear optimization formulation. In other works on robust gate sizing [37]–[40], the central idea is to capture the delay distributions by performing a statistical static timing analysis (SSTA), as opposed to the traditional STA, and then use either a general nonlinear programming technique or statistical sensitivity-based heuristic procedures to size the gates. In other work [41], the mean and variances of the node delays in the circuit graph are minimized in the selected paths, subject to constraints on delay and area penalty.

More formal optimization approaches have also been used. Approaches for optimizing the statistical power of the circuit, subject to timing yield constraints, can be presented as a convex formulation, as a second-order conic program [42]. For the binning model, a yield optimization problem is formulated [43], providing a binning yield loss function that has a linear penalty for delay of the circuit exceeding the target delay; the formulation is shown to be convex.

A gate sizing technique based on robust optimization theory has also been proposed [19], [44]: robust constraints are added to the original constraints set by modeling the intra-chip random process parameter variations as Gaussian variables, contained in a constant probability density uncertainty ellipsoid, centered at the nominal values.

A key problem in circuit optimization is the determination of sensitivities and criticality. This has also been the focus of considerable research [45]–[47].

D. Post-silicon sensor measurements

SSTA is a presilicon analysis technique used to determine the range of performance (delay or power) variations over a large population of dies. A complementary role, after the chip is manufactured, is played by post-silicon diagnosis, which is typically directed toward determining the performance of an individual fabricated chip based on measurements on that specific chip. This procedure provides particular information that can be used to perform post-silicon optimizations to make a fabricated part meet its specifications. Because presilicon analysis has to be generally applicable to the entire population of manufactured chips, the statistical analysis that it provides shows a relatively large standard deviation for the delay. On the other hand, post-silicon procedures, which are tailored to individual chips, can be expected to provide more specific information. Since tester time is generally prohibitively expensive, it is necessary to derive the maximum possible information through the fewest post-silicon measurements.

This issue has been addressed in several ways. In [48], post-silicon measurements are used to learn a more accurate spatial correlation model, which is fed back to the analysis stage to refine the statistical timing analysis framework. In [49], a path-based methodology is used for correlating post-silicon test data to presilicon timing analysis. In [50], a statistical gate sizing approach is studied to optimize the binning yield. Post-silicon debug methods and their interactions with circuit design are discussed in [51].

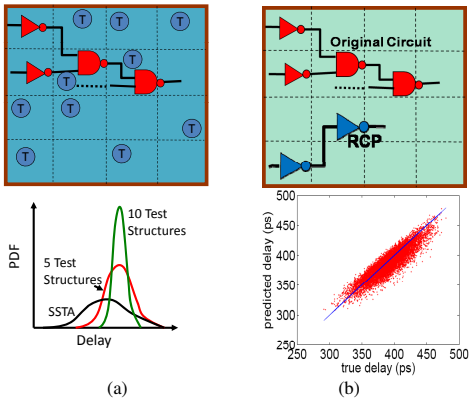

Fig. 5. Two alternative methods for placing test structures for fast post-silicon delay estimation: (a) inserting ring oscillators whose speed distribution can be used to infer ciruit delay characteristis (b) creating a representative critial path whose delay is highly correlated with the circuit delay.

Other approaches use presilicon SSTA analysis to guide postsilicon measurements. These methods use the results of a limited measurement are used to diagnose the performance of the manufactured part. A fast measurement-based approach [52] infers information from measurements from a small number of on-chip ring oscillators (ROs). These ROs are distributed over the area of the chip and are capable of capturing the variations of spatially correlated parameters over the die. The SSTA-based delay of the circuit is a Gaussian distribution, as are the SSTA-based delays of the ring oscillators, and these are correlated because of spatial correlations. If the ring oscillator delays are measured on a specific part, this part-specific information allows the delay of the part to be written as a conditional probability. By using enough stages to drown out the uncorrelated variations, the on-chip variations can be predicted on the part. This idea is illustrated in Figure 5(a), where the upper part shows the chip layout and the test structures, T, and the lower part shows how the variance of the conditional delay PDF can be narrowed down by adding more test structures: in other words, with more test structures, the delay can be predicted with greater confidence.

An alternative method [53] synthesizes a representative critical path (RCP) whose behavior tracks the worst-case delay of the circuit. The delays of both the original circuit and an

RCP can be computed from SSTA and are taken to be Gaussians; the RCP construction procedure explores the discrete solution space of the cells in the library and the placement (which determine spatial correlation effects) to maximize the correlation between the delay of the RCP and the original circuit. Figure 5(b) illustrates this idea, and shows a small RCP structure whose delay is correlated with a larger original circuit. The lower part of this figure shows a scatter plot illustrating the small mismatch between the delay predicted by the RCP and the actual circuit delay. Several such RCPs could be constructed and placed in the circuit, and a small number of measurements could yield a good estimate of the circuit delay.

III. Environmental variations

Unlike process variations, which are one-time variations that are static after a circuit is manufactured, environmental variations correspond to changes during the operation of a circuit. For process variations, it is possible to use statistical methods that optimize the manufacturing yield of the circuit, discarding (or binning) any die that fail to meet specifications. However, for any run-time environmental variations, it is essential to ensure that a circuit meets its specifications at all times. Therefore, environmental variations are subject to worst-case analysis, although approaches using statistical methods such as extreme value theory [54] have been proposed to identify the worst case. In the discussion below, we overview three types of environmental variations: due to supply voltage fluctuations, temperature changes, and soft errors.

A. Supply voltage variations

Run-time fluctuations in the supply voltage levels in a chip can cause significant variations in parameters such a gate delays, and may even result in logic failures. In nanometer-scale technologies, the current densities have increased over previous generations, and spatial imbalances between the currents in various parts of a chip are accentuated, particularly with the advent of multicore systems where some cores may switch on and off entirely. The drops along the supply and ground networks include IR drops due to large currents flowing in wires with nonzero resistances, as well as L dI/dt effects due to inductance.

Even for 2DICs, trends from the International Technology Roadmap for Semiconductors (ITRS) [55] indicate that the current delivered per power pin increases as technologies scale down, implying the likelihood of larger IR drops and L dI/dt noise. This effect becomes even more acute in 3DICs, where the amount of circuitry in a package increases, without a corresponding increase in the number of pins available to feed the increased current requirements. Therefore, it is important to analyze and optimize power grids to ensure correct circuit functionality and to minimize performance drifts.

1) Analysis: The analysis of power grids requires the solution of large RLC networks (that represent the interconnects in the power grid) with current sources (that model the functional blocks that draw current from the network) and voltage sources (that correspond to the V_{dd} source(s)). In general, the analysis problem corresponds to the solution of a set of modified nodal analysis equations of the form:

$$C\frac{d\mathbf{V}(t)}{dt} + G\mathbf{V}(t) = \mathbf{J}(t) \tag{2}$$

where $\mathbf{V}$ is the vector of unknowns, G is the conductance matrix, C captures the capacitance and inductance terms, and $\mathbf{J}$ is the vector of current excitations to the system. The system of equations to be solved is large, typically involving millions of variables. This system of equations is typically sparse and positive definite, but its large dimension necessitates the use of efficiency-enhancing methods. Specifically:

- *Hierarchical methods* [56] may use either natural hierarchies or specified hierarchies to solve the problem efficiently. Blocks in lower levels of the hierarchy are represented using sparse macromodels, corresponding to sparsified Schur complements. The global grid, along with these macromodels is then solved, and these solutions are propagated to the local grids. This approach leads to large savings in computation time and memory usage.
- *Multigrid methods* [57]–[59] successively coarsen the grid by reducing the number of nodes in the network. The coarsened grid is solved to obtain an approximate solution that captures the low-frequency spatial components of the voltage variation. This solution is then transformed back to the original grid through interpolation operators, capturing high-frequency spatial components. Through multiple iterations of the so-called "V-cycle" [60], further accuracy is achievable.
- *Random walk methods* leverage an analogy [61] between random walks and power grids to solve the network [62], [63]. These methods are particularly useful for local and incremental solves [64], but may also be used for approximate full solves, or to generate preconditioners for more exact solves [65].

Power grid analysis is typically performed under two scenarios:

- DC analysis solutions are useful in early stages of design.
- Transient analysis solutions are necessary for more detailed analyses later in the design process.

Transient solutions may be computed either using time-stepping (constant time steps are typically used) or using model order reduction methods. Time-domain techniques are popularly used in many tools, and several techniques for solving the analysis problem have been proposed.

2) Optimization: While analysis techniques can diagnose problems in a power grid, it is essential to build optimization techniques that can correct these problems and build reliable power grids. Effective techniques for optimization include pin assignment [66], [67], topology optimization [68], [69], wire sizing [70], [71], and decoupling capacitor (decap) insertion [72]. The last of these deliberately inserts capacitors into the power grid: these act as charge reservoirs that damp down the effects of fast transients by providing a nearby source of charge to feed the current drawn by the functional blocks. As on-chip capacitors grow more leaky, however, further enhancements are required in decap allocation. Recent research has led to the

possibility of using MIM capacitors (or trench capacitors) as decaps [73], and also to the notion of novel low-leakage decap technologies [74].

B. Thermal variations

The impact of temperature on the functioning of a chip is an important factor in inducing variation and reliability issues. Elevated on-chip temperatures can have several consequences on performance. First, they cause transistors threshold voltages to go down, and carrier mobilities to increase: the former tends to speed up a circuit, while the latter tends to slow it down. Depending on which effect wins, a circuit may show either negative temperature dependence if the delay increases with temperature, positive temperature dependence if it decreases with temperature, or mixed temperature dependence if the trend is nonuniform. Second, leakage power increases with temperature: in cases where this increase is substantial, the increased power can raise the temperature further, causing a feedback cycle. This positive feedback can even cause thermal runaway, where the increase in the power goes to a point that cannot be supported by the heat sink, and the chip burns out. Third, reliability effects, such as bias temperature instability and electromigration generally degrade with temperature, implying that higher temperatures tend to age a circuit faster. In 3DIC technologies, disparities between the coefficients of thermal expansion of through-silicon vias (TSVs) and the surrounding silicon can result in fatigues or cracks, and in altered transistor mobilities in a region surrounding the TSV.

1) Analysis: Conventional heat transfer on a chip is described by Fourier's law of conduction. More fine-grained nanoscale thermal analysis can be performed by modeling electron-phonon interactions, involving the solution of the Boltzmann transport equation [75], but Fourier-based models are adequate for full-chip analysis. This analysis requires the solution of the partial differential equation:

$$\rho c_p \frac{\partial T(\mathbf{r},t)}{\partial t} = k_t \nabla^2 T(\mathbf{r},t) + g(\mathbf{r},t) \tag{3}$$

where ρ is the density of the material, c_p is its heat capacity, T is the temperature, k_t is the thermal conductivity of the material, and g is the power density per unit volume. The boundary conditions for this equation are typically described in Dirichlet form, specifying information related to heat sinks that lie at the boundary of the chip.

Like the analysis of power grids, thermal analysis can be performed for the DC case (where the left-hand side of the above equation becomes zero and the equation reduces to a Poisson's equation) or for the transient case. The time constants of heat transfer are much longer than the clock period for today's VLSI circuits, and if a circuit remains within the same power mode for an extended period of time, and its power density distribution remains relatively constant, steady-state analysis can capture the thermal behavior of the circuit accurately. Even if this is not the case, steady-state analysis can be particularly useful for early and more approximate analysis, in the same spirit that steady-state analysis is used to analyze power grid networks early in the design cycle. On the other hand, when greater levels of detail about the inputs are available, transient analysis is possible and potentially useful.

The similarities with power grids go further: under finite difference discretization, thermal analysis can be shown to be equivalent to solving an RC circuit with current and voltage sources [76], leading to an equation similar to (2). The equation corresponds to a network where "thermal resistors" are connected between nodes that correspond to spatially adjacent regions, "thermal capacitors" to ground, and "thermal current sources" that map on to power sources. The voltages at the nodes in this thermal circuit can then be computed by solving this circuit, and these yield the temperature at each node. This implies that similar solution methods may be employed for thermal analysis as for power grid analysis.

However, there are also some differences between the problems. The power sources (or current sources under the duality) in thermal analysis lie on one specific layer of the grid for 2D circuits, or in a discrete set of layers for 3DICs. This may be leveraged to perform Green-function-based analysis [77], [78]. Moreover, the symmetry in thermal conductivities lends itself to the use of fast Poisson solver (FPS) methods [79], and a relationship between Green functions and FPS methods is demonstrated in [80].

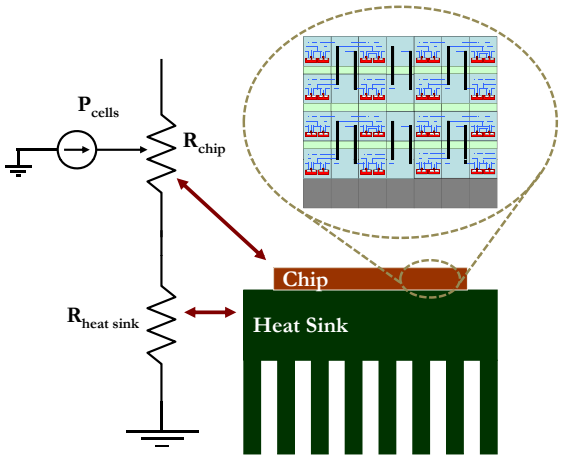

Fig. 6. A schematic of the thermal environment for a 3DIC, illustrating the considerable scope for thermal optimization.

2) Optimization: Figure 6 shows a schematic of 3DIC with the associated packaging. At left, we see a simplified thermal model, with a distributed current source representing the on-chip power dissipation, a distributed resistive network showing the on-chip thermal resistance, and a sink resistance corresponding to the package, connected to the ambient (ground node). We address the DC case in this example, but a similar set of conclusions may be drawn for the transient case.

The temperature in the system corresponds to the voltage at a node in the circuit, and thermal optimization corresponds to the problem of designing the system so as to reduce this voltage. This may be achieved by one of several means:
Reducing the value of the current source: Since current in the thermal circuit corresponds to power, this essentially implies that low-power circuits are more likely to be thermally op-

timized. Such optimizations may be performed at all levels, ranging from microarchitectural optimizations [81]–[85] to circuit optimizations.

Reducing the voltage in the distributed thermal resistance: This may be achieved by appropriately managing the spatial distribution of power sources within the thermal network, employing design techniques such as floorplanning [86]–[88] and placement [89]–[93]. Another technique involves the reduction of on-chip thermal resistance using active or passive on-chip cooling techniques. Passive cooling methods include the use of thermal vias, which are TSVs whose function is to conduct heat from the warm areas towards the heat sink and ambient, and such optimizations are discussed in [94]–[96].

Reducing the voltage across the sink resistor: This corresponds to using a better heat sink, with a lower thermal resistance. However, the cost of a heat sink, beyond a point, increases very steeply with the desired reductions in the thermal resistance: for example, as we approach the limits of air cooling towards liquid cooling, the cost of the cooling solution is greatly increased.

Mitigation techniques: For regions where the performance is significantly degraded due to thermal effects and the temperature cannot be reduced, mitigation techniques can be used to overcome the degradation in performance. Such approaches include the use of adaptive body biases, adaptive supply voltages, and frequencies [97]–[104]. Such approaches may be facilitated by the use of efficient timing analysis methods in the presence of body bias, such as those presented in [105].

C. Soft errors

With the number of devices on a chip numbering in the billions, and with limited charge storage ability for each device, integrated circuits are increasingly susceptible to strikes from cosmic rays, alpha particles, and neutron-induced Boron fission [106]. These strikes can cause momentary surges in charge that can result in effects such as increased delays, logic failures, or incorrectly flipped memory bits. These impermanent errors are referred to as soft errors, and these have been observed to be significant, not only in radiation-sensitive environments such as space, but also in normal high-performance applications on earth. Not every single-event upset may result in incorrect logic: in particular, mechanisms such as logical masking, temporal masking, and electrical masking [107] can render such events harmless in digital logic. However, the problem is serious enough to merit significant research efforts.

Aside from the use of error-correcting codes in memories, methods for radiation-hardening include special process techniques that add guard-bands around devices, techniques such as gate sizing and threshold voltage assignment [108] that strengthen pull-up/pull-down devices and retain charged states in a gate, and special layout techniques [109] engineered to improve soft error resilience.

IV. Aging Mechanisms

A. Bias temperature instability

Bias temperature instability is a phenomenon that causes threshold voltage shifts over long periods of time, eventually causing the circuit to fail to meet its specifications. The word "bias" refers to the fact that this degradation is heightened by the application of a bias on the gate node of a transistor.

The phenomenon of negative bias temperature instability (NBTI) can be illustrated with the help of a simple circuit, an inverter, illustrated in Figure 7(a). When a PMOS transistor is biased in inversion ($V_{gs} = -V_{dd}$) (for example, when the input of the inverter is at logic 0), interface traps are generated due to the dissociation of $Si - H$ bonds along the substrate-oxide interface, as illustrated in Figure 7(b). The connection of this mechanism to thermal effects is that the rate of generation of these traps is accelerated by elevated temperatures, and therefore, increased on-chip temperatures can directly affect the lifetime of a chip. The time for which the transistor is stressed is another factor that increases the level of degradation. These traps cause an increase in the threshold voltage (V_{th}), and a reduction in the saturation current (I_{dsat}) of the PMOS transistors. This effect, known as NBTI, has become a significant reliability issue in high-performance digital IC design, especially in sub-130nm technologies [110]–[115]. An increase in V_{th} causes the circuit delay to degrade, and when this degradation exceeds a certain amount, the circuit may fail to meet its timing specifications. The rate constants of the reactions that define NBTI are dependent on temperature, and are worsened at elevated temperatures.

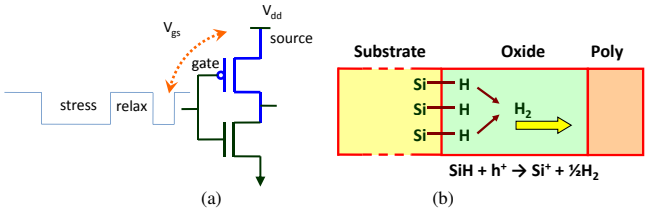

Fig. 7. (a) An inverter whose PMOS device is alternately subjected to NBTI stress and relax phases (b) An illustration of the phenomenon of NBTI

A corresponding and dual effect, known as Positive Bias Temperature Instability (PBTI) can be seen for NMOS devices, for example, when the input to an inverter is at logic 1, and a positive bias stress is applied across the gate oxide of the NMOS device. Although the impact of a stressing bias on PBTI is lower than NBTI [116], PBTI is becoming increasingly important in its own right. Moreover, techniques are developed to reduce NBTI can contribute to increasing PBTI. For the example of the inverter listed earlier, if the input is biased so that it is more likely to be at logic 1 than logic 0, the NBTI stress on the PMOS device is reduced since the V_{gs} bias becomes zero; however, this now places a bias on the NMOS device, which now has a nonzero V_{gs} value.

At the circuit level, the effect of bias temperature instability (BTI) is in alterations of the transistor threshold voltages. Under DC stress, the threshold voltage of a PMOS transistor degrades with time, t, at a rate given by

$$\Delta V_{th} \propto t^{1/6} \tag{4}$$

However, in general, transistors in a circuit are not continuously stressed, but a sequence of alternating 0s and 1s is applied at

their gate nodes. When the stress is removed, the threshold voltage is seen to recover towards its original value. Analytical models for the change in threshold voltage are provided in [113], [117]. For NBTI, the degradation in the threshold voltage is proportional to the probability p that the signal at the gate of the transistor is at logic 0, corresponding to the proportion of time that the transistor is likely to be under stress. A similar model for PBTI concludes, analogously, that the degradation is proportional to $1 - p$.

Several techniques have been presented in the literature for NBTI optimization. For digital logic, the problem may be resolved prior to manufacturing by appropriately padding the delay specifications to allow for degradation over the life of the circuit. Static methods proposed for addressing this include gate sizing, resynthesis, and technology mapping [118], which not only adjust the timing but also have the capability of changing signal probabilities.

Static methods incur high overheads due to the padded delays: instead, dynamic run-time methods may be used with a small static overhead. Moreover, it can be observed that over the life of the circuit, the delay increases but the leakage power decreases. As a result of the latter, the circuit may be well within its power budget late in its life, implying that one must overdesign the resources used to control power and temperature based on the requirements at the beginning of life.

The approach in [119] attempts to dynamically adjust the circuit delay using a combination of adaptive body biases and adaptive supply voltages. This method begins with a smaller initial delay padding than the static method, and alters the body biases and supply voltages over time using a sensor-based method, ensuring that the timing specification is met, while staying close to the power budget throughout the circuit lifetime. In this case, a time sensor is used, but silicon odometer methods [120] could also be adapted to similar schemes.

Figure 8 shows the change in the delay and power as a function of time for five scenarios. The nominal case, which meets the delay specifications, violates the delay constraint almost immediately and is not viable. The static padding case uses delay padding to meet the timing constraints at the end of life, and incurs significant power overheads, even accounting for the fact that the leakage reduces with time. The adaptive approach uses adaptive body biases and adaptive Vdd values; the slight initial padding is due to the discreteness of the cell library. The hybrid approach, which is the best solution, using a mix of initial delay padding and adaptive behavior. The sum of dynamic and leakage power is seen to show the best overall trend.

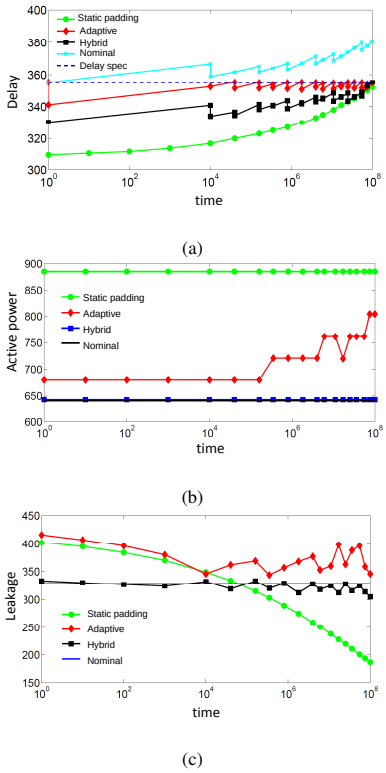

Fig. 8. The change in the (a) delay, (b) active power, and (c) leakage power of the benchmark "des" using various optimizations.

B. Gate oxide breakdown

Time-dependent dielectric breakdown (TDDB) in the gate oxide is a reliability phenomenon in gate oxides that results in a sudden discontinuous increase in the conductance of the gate oxide at the point of breakdown, as a result of which the current through the gate insulator increases significantly. This phenomenon, illustrated in Figure 9(a), is of particular concern as gate oxide thicknesses become thinner with technology scaling, and gates become more susceptible to breakdown. Various models for explaining TDDB have been put forth, including the hydrogen model, the anode-hole injection model, the thermochemical model (also known as the E model, where E is the electric field across the oxide), and the percolation model: for a survey, the reader is referred to [121], [122]. Unlike BTI, this mechanism is not known to be reversible, and any damage caused can be assumed to be permanent.

The time to breakdown, T_{BD}, can be modeled statistically using a Weibull distribution, whose cumulative density function (CDF) is given by

$$CDF(T_{BD}) = 1 - exp\left(\left[-\left(\frac{T_{BD}}{\alpha}\right)^{\beta}\right]\right) \quad (5)$$

The parameter α corresponds to the time-to-breakdown at about the 63rd percentile, and β is the Weibull slope.

At the circuit level, the traditional failure prediction method for a large circuit uses area-scaling, extrapolated from single device characterization [123]. The idea is based on the weakest-link assumption, that the failure of any individual device will cause the failure of the whole chip. Recently, new approaches have been proposed to improve the prediction accuracy by empirical calibration using real circuit test data [124], or by considering the variation of gate-oxide thickness [125]. The former is empirical and hard to generalize, while the latter does not consider the effect of breakdown location. Moreover, all existing methods circuit-level methods assume that (a) the transistors in the circuit are *always* under stress, and (b) any transistor breakdown *always* leads to a circuit failure.

In general, the above assumptions are not true. The work in

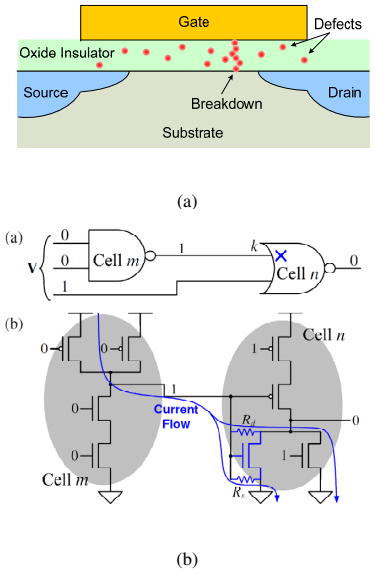

Fig. 9. (a) The phenomenon of gate oxide breakdown. (b) An example that illustrates the inherent resilience of CMOS circuits to oxide breakdown failures.

[126] develops a scalable method for analyzing the catastrophic failure probability of large digital circuits, while incorporating inherent circuit resilience to logic failure as follows:

- At the transistor level, it captures the fact that only PMOS transistors in inversion mode are susceptible to hard breakdowns, which cause logic failures, and adjust the Weibull time-to-breakdown model to incorporate the actual stressing modes of NMOS transistors in the inversion mode.
- At the logic cell level, it recognizes that the leakage current due to a breakdown event leads to a resistive divider, and depending on the breakdown resistance (which is a function of the location of the breakdown in the transistor) and the strength of the opposing transistor, the breakdown may or may not lead to a logic failure. This is illustrated in Figure 9(b), where the breakdown in cell n causes breakdown current; depending on the strength of the opposing transistors, this may (or may not) change the logic valud at the output(s) of cell(s) m and n.
- At the circuit level, it derives a closed-form expression for the failure probability of a large circuit, and demonstrates that this is in the form of a Weibull distribution.

The results of this approach indicate that the area scaling model can be between half and one order of magnitude off in predicting the circuit lifetime. Moreover, the approach points to ways of enhancing the resilience of a circuit to oxide breakdown: by sizing up transistors to favor logic value retention in the event of breakdown. A geometric program formulation is provided in [126] to capture this optimization.

Gate oxide breakdown leads to alterations in the DC noise margins for read, write, and retention, as well as in the read and write access times. Methods for enhancing the reliability of memory have also been studied [127], and monitoring schemes have been proposed to enhance memory reliability.

C. Hot carrier injection

Degradation due to hot carrier injection (HCI) appears mainly in NMOS devices when a large drain-to-source voltage and gate-to-source voltage is applied. HCI manifests itself as an increase in threshold voltage and a decline in channel mobility, leading to degradation in transistor drain current [128]–[130]. HCI is caused by various effects: the traditional explanation was based on impact ionization, but more complex effects are seen in nanometer-scale technologies. With supply voltages leveling off even as geometries shrink, HCI will worsen in future technologies, and is likely to be the dominant effect for long-term failures.

D. Interconnect-related reliability issues

The effect of aging and thermal effects can significantly impact interconnect integrity. Two examples of such issues are related to electromigration and TSV-induced variations in 3DICs.

The phenomenon of electromigration is related to the effects of current patterns applied to a wire over a long period of time, due to which atoms in the wire are seen to physically migration, particularly in regions where the current density is high. This can cause the wire to have increased resistance as it is thinned, or even become an open-circuit, and is therefore a serious reliability problem. This problem is witnessed most notably in supply (power and ground) wires [131], [132], where the flow of current is mostly unidirectional, but AC electromigration is also seen in signal wires [133]. Amelioration strategies for electromigration are primarily built in by ensuring that the current density on a wire never exceeds a specified threshold. For the same current, the use of wider wires results in lower current densities, and therefore, wire-widening is a potent tool for overcoming electromigration, with its accompanying overheads in taking up additional routing area and potentially on signal lines, increased power.

The mean time to failure (MTTF) of a wire under electromigration is described by Black's equation:

$$MTTF = AJ^{-n}e^{Q/kT} \tag{6}$$

where J is the average current density in the wire, n is an empirical exponent whose value is about 2, Q is the activation energy for grain-boundary diffusion, equal to about 0.7eV for Al-Cu, k is Boltzmann's constant, and T is the temperature of the wire.

Traditionally, electromigration has been controlled by limiting the value of the current density, J, in a wire. This is imposed as an extra constraint, in addition to IR drop and L dI/dt constraints during power grid optimization.

Variations due to TSVs are caused by the uneven coefficient of thermal expansion (CTE) between the metal in the TSV and the surrounding silicon. The variation causes significant stress, which in turn causes the mobility of devices to vary within a radius of the TSV. Analyses of TSV-induced stress are presented in, for example, [134], and their effects on on-chip design are explored in [135], [136]. One way of overcoming this CTE mismatch is to use tungsten TSVs, since tungsten has a very similar CTE as silicon, instead of copper. However,

this involves considerable sacrifice in electrical properties since tungsten is also significantly less conductive than copper, implying the need for larger TSVs.

V. Conclusion

This paper has attempted to provide an overview of the underlying causes and effects of on-chip variations, and to provide a snapshot of research progress and needs in the area of design to overcome such variations. As circuit technologies continue to scale, the need for such design techniques will become increasingly important and novel techniques that go well beyond the ideas described in this paper must be developed.

References

[1] S. Borkar, "Design and reliability challenges in nanometer technologies," in *Proceedings of the ACM/IEEE Design Automation Conference*, 2004.

[2] K. Kuhn, C. Kenyon, A. Kornfeld, M. Liu, A. Maheshwari, W. k. Shih, S. Sivakumar, G. Taylor, P. VanDerVoorn, and K. Zawadzki, "Managing process variation in Intels 45nm CMOS technology," *Intel Technology Journal*, vol. 12, June 2008.

[3] H. S. Momose, S.-I. Nakamura, T. Ohguro, T. Yoshitomi, E. Morifuji, T. Morimoto, Y. Katsumata, and H. Iwai, "Study of the manufacturing feasibility of 1.5-nm direct-tunneling gate oxide MOSFETs: uniformity, reliability, and dopant penetration of the gate oxide," *IEEE Transactions on Electron Devices*, vol. 45, pp. 691–700, Mar. 1998.

[4] M. J. M. Pelgrom, A. C. J. Duinmaijer, and A. P. G. Welbers, "Matching properties of MOS transistors," *IEEE Journal of Solid-State Circuits*, vol. 24, pp. 1433–1440, Oct. 1989.

[5] H. Chang and S. S. Sapatnekar, "Statistical timing analysis considering spatial correlations using a single PERT-like traversal," in *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*, pp. 621–625, 2003.

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

[7] R. Chen, L. Zhang, V. Zolotov, C. Visweswariah, and J. Xiong, "Static timing: back to our roots," in *Proceedings of the Asia/South Pacific Design Automation Conference*, pp. 310–315, Jan. 2008.

[8] S. Bhardwaj, S. Vrudhula, P. Ghanta, and Y. Cao, "Modeling of intra-die process variations for accurate analysis and optimization of nano-scale circuits," in *Proceedings of the ACM/IEEE Design Automation Conference*, pp. 791–796, July 2006.

[9] J. Xiong, V. Zolotov, and L. He, "Robust extraction of spatial correlation," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 26, pp. 619–631, Apr. 2007.

[10] F. Liu, "A general framework for spatial correlation modeling in VLSI design," in *Proceedings of the ACM/IEEE Design Automation Conference*, (San Diego, CA), pp. 817–822, 2007.

[11] S. S. Sapatnekar, *Timing*. Boston, Massachusetts, USA: Springer, 2004.

[12] D. Morrison, *Multivariate Statistical Methods*. New York, NY: McGraw-Hill, 1976.

[13] H. Chang and S. S. Sapatnekar, "Statistical timing analysis under spatial correlations," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 24, pp. 1467–1482, Sept. 2005.

[14] C. Clark, "The greatest of a finite set of random variables," *Operations Research*, vol. 9, pp. 85–91, March–April 1961.

[15] X. Li, J. Le, P. Gopalakrishnan, and L. T. Pileggi, "Asymptotic probability extraction for non-normal distributions of circuit performance," in *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*, pp. 2–9, 2004.

[16] X. Li, J. Le, P. Gopalakrishnan, and L. T. Pileggi, "Asymptotic probability extraction for nonnormal performance distributions," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 26, pp. 16–37, Jan. 2007.

[17] Y. Zhan, A. Strojwas, X. Li, and L. Pileggi, "Correlation-aware statistical timing analysis with non-Gaussian delay distributions," in *Proceedings of the ACM/IEEE Design Automation Conference*, pp. 77–82, 2005.

[18] J. Singh and S. S. Sapatnekar, "Statistical timing analysis with correlated non-Gaussian parameters using independent component analysis," in *Proceedings of the ACM/IEEE Design Automation Conference*, pp. 155–160, 2006.

[19] J. Singh and S. S. Sapatnekar, "A scalable statistical static timing analyzer incorporating correlated non-Gaussian and Gaussian parameter variations," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 27, pp. 160–173, Jan. 2008.

[20] A. Hyvärinen and E. Oja, "Independent component analysis: Algorithms and applications," *Neural Networks*, vol. 13, pp. 411–430, 2000.

[21] H. Chang, V. Zolotov, S. Narayan, and C. Visweswariah, "Parameterized block-based statistical timing analysis with non-Gaussian parameters, nonlinear delay functions," in *Proceedings of the ACM/IEEE Design Automation Conference*, pp. 71–76, 2005.

[22] V. Khandelwal and A. Srivastava, "A general framework for accurate statistical timing analysis considering correlations," in *Proceedings of the ACM/IEEE Design Automation Conference*, pp. 89–94, 2005.

[23] V. Khandelwal and A. Srivastava, "A quadratic modeling-based framework for accurate statistical timing analysis considering correlations," *IEEE Transactions on VLSI Systems*, vol. 15, pp. 206–215, Feb. 2007.

[24] L. Cheng, J. Xiong, and L. He, "Non-linear statistical static timing analysis for non-Gaussian variation sources," in *Proceedings of the ACM/IEEE Design Automation Conference*, pp. 250–255, 2007.

[25] F. N. Najm, N. Menezes, and I. A. Ferzli, "A yield model for integrated circuits and its application to statistical timing analysis," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 26, pp. 574–591, Mar. 2007.

[26] S. Onaissi and F. N. Najm, "A linear-time approach for static timing analysis covering all process corners," in *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*, pp. 217–224, 2006.

[27] S. V. Kumar, C. V. Kashyap, and S. S. Sapatnekar, "A framework for block-based timing sensitivity analysis," in *Proceedings of the ACM/IEEE Design Automation Conference*, 2008.

[28] D. Lee, W. Kwong, D. Blaauw, and D. Sylvester, "Analysis and minimization techniques for total leakage considering gate oxide leakage," in *Proceedings of the ACM/IEEE Design Automation Conference*, pp. 175–180, June 2003.

[29] R. Rao, A. Srivastava, D. Blaauw, and D. Sylvester, "Statistical estimation of leakage current considering inter- and intra-die process variation," in *Proceedings of the International Symposium of Low Power Electronic Devices*, pp. 84–89, Aug. 2003.

[30] R. Rao, A. Devgan, D. Blaauw, and D. Sylvester, "Parametric yield estimation considering leakage variability," in *Proceedings of the ACM/IEEE Design Automation Conference*, pp. 442 – 447, June 2004.

[31] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*. New York, NY: McGraw-Hill, 3rd ed., 1991.

[32] A. A. Abu-Dayya and N. C. Beaulieu, "Comparison of methods of computing correlated lognormal sum distributions and outages for digital wireless applications," in *IEEE 44th Vehicular Technology Conference, vol. 1*, pp. 175–179, June 1994.

[33] H. Chang and S. S. Sapatnekar, "Full-chip analysis of leakage power under process variations, including spatial correlations," in *Proceedings of the ACM/IEEE Design Automation Conference*, (Anaheim, CA), pp. 523–528, June 2005.

[34] A. Srivastava, S. Shah, K. Agarwal, D. Sylvester, D. Blaauw, and S. W. Director, "Accurate and efficient gate-level parametric yield estimation considering correlated variations in leakage power and performance," in *Proceedings of the ACM/IEEE Design Automation Conference*, pp. 535–540, June 2005.

[35] H. Chang and S. S. Sapatnekar, "Prediction of leakage power under process uncertainties," *ACM Transactions on Design Automation of Electronic Systems*, vol. 12, Apr. 2007. Article 12 (27 pages).

[36] E. Jacobs and M. Berkelaar, "Gate sizing using a statistical delay model," in *Proceedings of Design, Automation, and Test in Europe*, pp. 283–290, 2000.

[37] S. H. Choi, B. C. Paul, and K. Roy, "Novel sizing algorithm for yield improvement under process variation in nanometer technology," in *Proceedings of the ACM/IEEE Design Automation Conference*, pp. 454–459, June 2004.

[38] D. Sinha, N. V. Shenoy, and H. Zhou, "Statistical gate sizing for timing yield optimization," in *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*, pp. 1037–1042, Nov. 2005.

[39] A. Agarwal, K. Chopra, D. Blaauw, and V. Zolotov, "Circuit optimization using statistical static timing analysis," in *Proceedings of the ACM/IEEE Design Automation Conference*, pp. 338–342, June 2005.

[40] K. Chopra, S. Shah, A. Srivastava, D. Blaauw, and D. Sylvester, "Parametric yield maximization using gate sizing based on efficient statistical power and delay gradient computation," in *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*, pp. 1023–1028, Nov. 2005.

[41] S. Raj, S. B. K. Vrudhala, and J. Wang, "A methodology to improve timing yield in the presence of process variations," in *Proceedings of the ACM/IEEE Design Automation Conference*, pp. 448–453, June 2004.

[42] M. Mani, A. Devgan, and M. Orshansky, "An efficient algorithm for statistical power under timing yield constraints," in *Proceedings of the ACM/IEEE Design Automation Conference*, pp. 309–314, June 2005.

[43] A. Davoodi and A. Srivastava, "Variability driven gate sizing for binning yield optimization," in *Proceedings of the ACM/IEEE Design Automation Conference*, pp. 959–964, 2006.

[44] J. Singh, V. Nookala, T. Luo, and S. Sapatnekar, "Robust gate sizing by geometric programming," in *Proceedings of the ACM/IEEE Design Automation Conference*, pp. 315–320, June 2005.

[45] X. Li, J. Le, M. Celik, and L. T. Pileggi, "Defining statistical sensitivity for timing optimization of logic circuits with large-scale process and environmental variations," in *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*, pp. 844–851, Nov. 2005.

[46] J. Xiong, V. Zolotov, N. Venkateswaran, and C. Visweswariah, "Criticality computation in parameterized statistical timing," in *Proceedings of the ACM/IEEE Design Automation Conference*, pp. 63–68, July 2006.

[47] H. Mogal, H. Qian, S. S. Sapatnekar, and K. Bazargan, "Clustering based pruning for statistical criticality computation under process variations," in *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*, pp. 340–343, Nov. 2007.

[48] B. Lee, L. Wang, and M. S. Abadir, "Refined statistical static timing analysis through learning spatial delay correlations," in *Proceedings of the ACM/IEEE Design Automation Conference*, pp. 149–154, July 2006.

[49] L. Wang, P. Bastani, and M. S. Abadir, "Design-silicon timing correlation–a data mining perspective," in *Proceedings of the ACM/IEEE Design Automation Conference*, pp. 385–389, June 2007.

[50] A. Davoodi and A. Srivastava, "Variability driven gate sizing for binning yield optimization," in *Proceedings of the ACM/IEEE Design Automation Conference*, pp. 956–964, July 2006.

[51] M. Abramovici, P. Bradley, K. Dwarakanath, P. Levin, G. Memmi, and D. Miller, "A reconfigurable design-for-debug infrastructure for SoCs," in *Proceedings of the ACM/IEEE Design Automation Conference*, pp. 7–12, July 2006.

[52] Q. Liu and S. S. Sapatnekar, "Confidence scalable post-silicon statistical delay prediction under process variations," in *Proceedings of the ACM/IEEE Design Automation Conference*, pp. 497–502, June 2007.

[53] Q. Liu and S. S. Sapatnekar, "Synthesizing a representative critical path for post-silicon delay prediction," in *Proceedings of the ACM International Symposium on Physical Design*, pp. 183–190, Apr. 2009.

[54] N. E. Evmorfopoulos, G. I. Stamoulis, and J. N. Avaritsiotis, "A Monte Carlo approach for maximum power estimation based on extreme value theory," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 21, pp. 415–432, Apr. 2002.

[55] Semiconductor Industry Association, *International Technology Roadmap for Semiconductors*, 1997-2005.

[56] M. Zhao, R. V. Panda, S. S. Sapatnekar, and D. Blaauw, "Hierarchical analysis of power distribution networks," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 21, pp. 159–168, Feb. 2002.

[57] J. Kozhaya, S. R. Nassif, and F. N. Najm, "A multigrid-like technique for power grid analysis," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 21, pp. 1148–1160, Oct. 2002.

[58] H. Su, E. Acar, and S. R. Nassif, "Power grid reduction based on algebraic multigrid principles," in *Proceedings of the ACM/IEEE Design Automation Conference*, pp. 109–112, 2003.

[59] K. Wang and M. Marek-Sadowska, "On-chip power-supply network optimization using multigrid-based technique," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 24, pp. 407–417, Mar. 2005.

[60] W. L. Briggs, "A multigrid tutorial." `http://www.llnl.gov/CASC/people/henson/mgtut/ps/mgtut.pdf`.

[61] P. G. Doyle and J. L. Snell, *Random Walks and Electric Networks*. Washington, DC, USA: Mathematical Association of America, 1984.

[62] H. Qian, S. R. Nassif, and S. S. Sapatnekar, "Random walks in a supply network," in *Proceedings of the ACM/IEEE Design Automation Conference*, pp. 93–98, 2003.

[63] H. Qian, S. R. Nassif, and S. S. Sapatnekar, "Power grid analysis using random walks," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 24, pp. 1204–1224, Aug. 2005.

[64] B. Boghrati and S. S. Sapatnekar, "Incremental solution of power grids using random walks," in *Proceedings of the Asia/South Pacific Design Automation Conference*, pp. 757–762, 2010.

[65] H. Qian and S. S. Sapatnekar, "A hybrid linear equation solver and its application in quadratic placement," in *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*, pp. 905–909, 2005.

[66] M. Zhao, Y. Fu, V. Zolotov, S. Sundareswaran, and R. Panda, "Optimal placement of power-supply pads and pins," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 25, pp. 144–154, Jan. 2006.

[67] T. Sato, H. Onodera, and M. Hashimoto, "Successive pad assignment algorithm to optimize number and location of power supply pad using incremental matrix inversion," in *Proceedings of the Asia/South Pacific Design Automation Conference*, pp. 723–728, 2005.

[68] J. Singh and S. S. Sapatnekar, "Congestion-aware topology optimization of structured power/ground networks," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 24, pp. 683–695, May 2005.

[69] J. Singh and S. S. Sapatnekar, "A partition-based algorithm for power grid design using locality," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 25, pp. 664–677, Apr. 2006.

[70] H. Su, J. Hu, S. R. Nassif, and S. S. Sapatnekar, "Congestion-driven codesign of power and signal networks," in *Proceedings of the ACM/IEEE Design Automation Conference*, pp. 477–480, 2002.

[71] H. Su, J. Hu, S. S. Sapatnekar, and S. R. Nassif, "A methodology for the simultaneous design of supply and signal networks," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 23, pp. 1614–1624, Dec. 2004.

[72] H. Su, S. S. Sapatnekar, and S. R. Nassif, "Optimal decoupling capacitor sizing and placement for standard cell layout designs," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 22, pp. 428–436, Apr. 2003.

[73] P. Zhou, K. Sridharan, and S. S. Sapatnekar, "Congestion-aware power grid optimization for 3D circuits using MIM and CMOS decoupling capacitors," in *Proceedings of the Asia/South Pacific Design Automation Conference*, pp. 179–184, 2010.

[74] J. Gu, R. Harjani, and C. Kim, "Design and implementation of active decoupling capacitor circuits for power supply regulation in digital ics," *IEEE Transactions on VLSI Systems*, pp. 292–301, Feb. 2009.

[75] E. Pop, S. Sinha, and K. E. Goodson, "Heat generation and transport in nanometer-scale transistors," *Proceedings of the IEEE*, vol. 94, pp. 1587–1601, Aug. 2006.

[76] M. N. Ozisik, *Finite Difference Methods in Heat Transfer*. New York, New York, USA: CRC Press, 1994.

[77] Y. Zhan and S. S. Sapatnekar, "High efficiency Green function-based thermal simulation algorithms," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 26, pp. 1661–1675, September 2007.

[78] B. Wang and P. Mazumder, "Accelerated chip-level thermal analysis using multilayer green's function," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 26, pp. 325–344, Feb. 2007.

[79] Y. Saad, *Iterative Methods for Sparse Linear Systems*. Philadelphia, PA: SIAM, 2 ed., 2003.

[80] H. Qian and S. S. Sapatnekar, "Fast Poisson solvers for thermal analysis," in *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*, pp. 698–702, 2010.

[81] Y. Han, I. Koren, and C. A. Moritz, "Temperature aware floorplanning," in *Second Workshop on Temperature-Aware Computing Systems*, 2005.

[82] Y. W. Wu, C.-L. Yang, P.-H. Yuh, and Y.-W. Chang, "Joint exploration of architectural and physical design spaces with thermal consideration," in *Proceedings of the International Symposium of Low Power Electronic Devices*, pp. 123–126, 2005.

[83] K. Sankaranarayanan, S. Velusamy, M. Stan, and K. Skadron, "A case for thermal-aware floorplanning at the microarchitectural level," *The Journal of Instruction-Level Parallelism*, vol. 8, Sept. 2005.

[84] M. Healy, M. Vittes, M. Ekpanyapong, C. Ballapuram, S. K. Lim, H.-H. S. Lee, and G. H. Loh, "Microarchitectural floorplanning under performance and thermal tradeoff," in *Proceedings of Design, Automation, and Test in Europe*, pp. 1–6, 2006.

[85] V. Nookala, D. J. Lilja, and S. S. Sapatnekar, "Temperature-aware floorplanning of microarchitecture blocks with IPC-power dependence

modeling and transient analysis," in *Proceedings of the International Symposium of Low Power Electronic Devices*, pp. 298–303, 2006.

[86] J. Cong, J. Wei, and Y. Zhang, "A thermal-driven floorplanning algorithm for 3D ICs," in *Proceedings of the ACM International Symposium on Physical Design*, pp. 306–313, 2004.

[87] E. Wong and S. K. Lim, "3D floorplanning with thermal vias," in *Proceedings of Design, Automation, and Test in Europe*, pp. 878–883, 2006.

[88] P. Zhou, Y. Ma, Z. Li, R. P. Dick, L. Shang, H. Zhou, X. Hong, and Q. Zhou, "3D-STAF: Scalable temperature and leakage aware floorplanning for three-dimensional integrated circuits," in *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*, pp. 590–597, 2007.

[89] C. H. Tsai and S. M. Kang, "Cell-level placement for improving substrate thermal distribution," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 19, pp. 253–266, Feburary 2000.

[90] G. Chen and S. S. Sapatnekar, "Partition-driven standard cell placement," in *Proceedings of the ACM International Symposium on Physical Design*, pp. 75–80, 2003.

[91] B. Goplen and S. S. Sapatnekar, "Efficient thermal placement of standard cells in 3D ICs using a force directed approach," in *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*, pp. 86–89, November 2003.

[92] B. Goplen and S. S. Sapatnekar, "Placement of 3D ICs with thermal and interlayer via considerations," in *Proceedings of the ACM/IEEE Design Automation Conference*, pp. 626–631, 2007.

[93] J. Cong, G. Luo, J. Wei, and Y. Zhang, "Thermal-aware 3D IC placement via transformation," in *Proceedings of the Asia/South Pacific Design Automation Conference*, pp. 780–785, 2007.

[94] B. Goplen and S. S. Sapatnekar, "Thermal via placement in 3D ICs," in *Proceedings of the ACM International Symposium on Physical Design*, pp. 167–174, 2005.

[95] J. Cong and Y. Zhang, "Thermal-driven multilevel routing for 3-D ICs," in *Proceedings of the Asia/South Pacific Design Automation Conference*, pp. 121–126, 2005.

[96] T. Zhang, Y. Zhan, and S. S. Sapatnekar, "Temperature-aware routing in 3D ICs," in *Proceedings of the Asia/South Pacific Design Automation Conference*, pp. 309–314, 2006.

[97] L. Yan, J. Luo, and N. K. Jha, "Joint dynamic voltage scaling and adaptive body biasing for heterogeneous distributed real-time embedded systems," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 24, pp. 1030–1041, July 2005.

[98] L. Yan, J. Luo, and N. K. Jha, "Combined dynamic voltage scaling and adaptive body biasing for heterogeneous distributed real-time embedded systems," in *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*, pp. 30–37, 2003.

[99] S. M. Martin, K. Flautner, T. Mudge, and D. Blaauw, "Combined dynamic voltage scaling and adaptive body biasing for lower power microprocessors under dynamic workloads," in *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*, pp. 721–725, 2002.

[100] S. Yajuan, W. Zuodong, and W. Shaojun, "Energy-aware supply and body biasing voltage scheduling algorithm," in *Proceedings of the International Conference on Solid State and Integrated Circuits Technology*, pp. 1956–1959, 2004.

[101] A. Andrei, M. Schmitz, P. Eles, Z. Peng, and B. M. Al-Hashimi, "Overhead-conscious voltage selection for dynamic and leakage energy reduction of time-constrained systems," in *Proceedings of Design, Automation, and Test in Europe*, pp. 518–523, 2004.

[102] T. Fischer, F. Anderson, B. Patella, and S. Naffziger, "A 90nm variable-frequency clock system for a power-managed Itanium®-family processor," in *Proceedings of the IEEE International Solid-State Circuits Conference*, pp. 294–299,599, 2005.

[103] C. Piorier, R. McGowen, C. Bostak, and S. Naffziger, "Power and temperature control on an Itanium®-family processor," in *Proceedings of the IEEE International Solid-State Circuits Conference*, pp. 304–305, 2005.

[104] R. McGowen, C. A. Poirier, C. Bostak, J. Ignowski, M. Millican, W. H. Parks, and S. Naffziger, "Power and temperature control on a 90-nm Itanium family processor," *IEEE Journal of Solid-State Circuits*, vol. 41, pp. 229–237, Jan. 2006.

[105] S. Gupta and S. S. Sapatnekar, "Current source modeling in the presence of body bias," in *Proceedings of the Asia/South Pacific Design Automation Conference*, pp. 199–204, 2010.

[106] R. C. Baumann, "Soft errors in advanced semiconductor devices – Part i: The three radiation sources," *IEEE Transactions on Devices and Materials Reliability*, vol. 1, pp. 17–22, Mar. 2001.

[107] T. Karnik, P. Hazucha, and J. Patel, "Characterization of soft errors caused by single event upsets in CMOS processes," *IEEE Transactions on Dependable and Secure Computing*, vol. 1, pp. 128–143, apr–jun 2004.

[108] M. Choudhury, Q. Zhou, and K. Mohanram, "Design optimization for single-event upset robustness using simultaneous dual-VDD and sizing techniques," in *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*, pp. 204–209, 2006.

[109] H. Lee, K. Lilja, M. Bounasser, P. Relangi, I. Linscott, U. Inan, and S. Mitra, "LEAP: Layout design through error-aware placement for soft-error resilient sequential cell design," in *Proceedings of the IEEE International Reliability Physics Symposium*, pp. 203–212, 2010.

[110] V. Reddy, A. T. Krishnan, A. Marshall, J. Rodriguez, S. Natarajan, T. Rost, and S. Krishnan, "Impact of negative bias temperature instability on digital circuit reliability," in *Proceedings of the IEEE International Reliability Physics Symposium*, pp. 248–253, 2002.

[111] A. T. Krishnan, V. Reddy, S. Chakravarthi, J. Rodriguez, S. John, and S. Krishnan, "NBTI impact on transistor and circuit: Models, mechanisms and scaling effects," in *Proceedings of the IEEE International Electronic Devices Meeting*, pp. 14.5.1–14.5.4, 2003.

[112] D. K. Schroder, "Negative bias temperature instability: Physics, materials, process, and circuit issues," 2005. available at `http://www.ewh.ieee.org/r5/denver/sscs/Presentations/2005.08.Schroder.pdf`.

[113] M. A. Alam, "A critical examination of the mechanics of dynamic NBTI for pMOSFETs," in *Proceedings of the IEEE International Electronic Devices Meeting*, pp. 14.4.1–14.4.4, 2003.

[114] S. Chakravarthi, A. T. Krishnan, V. Reddy, C. Machala, and S. Krishnan, "A comprehensive framework for predictive modeling of negative bias temperature instability," in *Proceedings of the IEEE International Reliability Physics Symposium*, pp. 273–282, 2004.

[115] J. G. Massey, "NBTI: What we know and what we need to know - A tutorial addressing the current understanding and challenges for the future," in *Proceedings of the IEEE International Integrated Reliability Workshop Final Report*, pp. 199–211, 2004.

[116] V. Reddy, A. T. Krishnan, A. Marshall, J. Rodriguez, S. Natarajan, T. Rost, and S. Krishnan, "Impact of negative bias temperature instability on digital circuit reliability," in *Proceedings of the IEEE International Reliability Physics Symposium*, pp. 248–254, Apr. 2002.

[117] S. V. Kumar, C. H. Kim, and S. S. Sapatnekar, "An analytical model for negative bias temperature instability (NBTI)," in *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*, pp. 493–496, 2006.

[118] S. V. Kumar, C. H. Kim, and S. S. Sapatnekar, "NBTI-aware synthesis of digital circuits," in *Proceedings of the ACM/IEEE Design Automation Conference*, pp. 370–375, 2007.

[119] S. V. Kumar, C. H. Kim, and S. S. Sapatnekar, "Adaptive techniques for overcoming performance degradation due to aging in cmos circuits," *IEEE Transactions on VLSI Systems*. (to appear); available at `http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5371864&tag=1`.

[120] T.-H. Kim, R. Persaud, and C. H. Kim, "Silicon odometer: An on-chip reliability monitor for measuring frequency degradation of digital circuits," *IEEE Journal of Solid-State Circuits*, vol. 43, pp. 874–880, Apr. 2008.

[121] J. H. Stathis, "Reliability limits for the gate insulator in CMOS technology," *IBM Journal of Research and Development*, vol. 46, pp. 265–286, March/May 2002.

[122] E. Y. Wu, E. J. Nowak, A. Vayshenker, W. L. Lai, and D. L. Harmon, "CMOS scaling beyond the 100-nm node with silicon-dioxide-based gate dielectrics," *IBM Journal of Research and Development*, vol. 46, pp. 287–298, March/May 2002.

[123] J. H. Stathis, "Physical and predictive models of ultrathin oxide reliability in CMOS devices and circuits," *IEEE Transactions on Devices and Materials Reliability*, vol. 1, pp. 43–59, Mar. 2001.

[124] Y. H. Lee, N. Mielke, M. Agostinelli, S. Gupta, R. Lu, and W. McMahon, "Prediction of logic product failure due to thin-gate oxide breakdown," in *Proceedings of the IEEE International Reliability Physics Symposium*, pp. 18–28, Mar. 2006.

[125] K. Chopra, C. Zhuo, D. Blaauw, and D. Sylvester, "A statistical approach for full-chip gate-oxide reliability analysis," in *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*, pp. 698–705, Nov. 2008.

[126] J. Fang and S. S. Sapatnekar, "Scalable methods for the analysis and optimization of gate oxide breakdown," in *Proceedings of the IEEE International Symposium on Quality Electronic Design*, pp. 638–645, 2010.
[127] F. Ahmed and L. Milor, "Reliable cache design with detection of gate oxide breakdown using BIST," in *Proceedings of the IEEE International Conference on Computer Design*, pp. 366–371, 2009.
[128] K. Chen, S. Saller, I. Groves, and D. Scott, "Reliability effects on MOS transistors due to hot-carrier injection," *IEEE Transactions on Electron Devices*, vol. 32, pp. 386–393, Feb. 1985.
[129] H. Kufluoglu, *MOSFET Degradation due to Negative Bias Temperature Instability (NBTI) and Hot Carrier Injection (HCI), and Its Implications for Reliability-Aware VLSI Design*. PhD thesis, Purdue University, West Lafayette, IN, 2007.
[130] T. Nigam, "Impact of transistor level degradation on product reliability," in *Proceedings of the IEEE Custom Integrated Circuits Conference*, pp. 431–438, 2009.
[131] J. R. Black, "Electromigration failure modes in aluminum metallization for semiconductor devices," *Proceedings of the IEEE*, vol. 57, pp. 1587–1594, Sept. 1969.
[132] F. M. d'Heurle, "Electromigration and failure in electronics: An introduction," *Proceedings of the IEEE*, vol. 59, pp. 1409–1418, Oct. 1971.
[133] L. M. Ting, J. S. May, W. R. Hunter, and J. W. McPherson, "AC electromigration characterization and modeling of multilayered interconnects," in *Proceedings of the IEEE International Reliability Physics Symposium*, pp. 311–316, 1993.
[134] K. H. Lu, X. Zhang, S.-K. Ryu, J. Im, R. Huang, and P. S. Ho, "Thermo-mechanical reliability of 3-D ICs containing through silicon vias," in *Proceedings of the Electronic Components and Technology Conference*, 2009.
[135] J.-S. Yang, K. Athikulwongse, Y.-J. Lee, S. K. Lim, and D. Z. Pan, "TSV stress aware timing analysis with applications to 3D-IC layout optimization," in *Proceedings of the ACM/IEEE Design Automation Conference*, pp. 803–806, 2010.
[136] K. Athikulwongse, A. Chakraborty, J.-S. Yang, D. Z. Pan, and S. K. Lim, "Stress-driven 3D-IC placement with TSV keep-out zone and regularity study," in *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*, pp. 669–674, 2010.

Sachin S. Sapatnekar (S'86 M'93 F'03) received the B.Tech. degree from the Indian Institute of Technology, Bombay in 1987, the M.S. degree from Syracuse University in 1989, and the Ph.D. degree from the University of Illinois at Urbana-Champaign in 1992. From 1992 to 1997, he was an assistant professor in the Department of Electrical and Computer Engineering at Iowa State University. He is currently the Robert and Marjorie Henle Chair in the Department of Electrical and Computer Engineering and a Distinguished McKnight University Professor at the University of Minnesota.

He is an author/editor of eight books and several papers in the area of performance and layout issues in VLSI circuits. He has held positions on the editorial board of the IEEE Transactions on VLSI Systems, and the IEEE Transactions on Circuits and Systems II, the IEEE Transactions on CAD, and has been a Guest Editor for the latter. He has served on the Technical Program Committee for various conferences, as Technical Program and General Chair for the Tau workshop and ISPD, and is currently the Vice Chair for DAC2009. He is a recipient of the NSF Career Award, six conference Best Paper awards, and the SRC Technical Excellence award.