IPSJ Transactions on System LSI Design Methodology Vol. 1 18–32 (Aug. 2008)

Invited Paper

Variability and Statistical Design

Sachin S. Sapatnekar[†1]

With each technology generation, the effects of on-chip variations are seen to more profoundly affect digital circuit behavior. These variations may arise from fluctuations attributed to the manufacturing process (e.g., drifts in channel length, oxide thickness, threshold voltage, or doping concentration), which affect the circuit yield, as well as variations in the environmental operating conditions (e.g., supply voltage or temperature) after the circuit is manufactured, which impact the performance of the design. These effects can cause unacceptable alterations in circuit performance parameters such as timing and power, and variation-tolerant design is imperative for next-generation designs. This paper overviews research in this area, describing methods for the analysis and optimization of statistical effects.

1. Introduction

Current-day integrated circuits are afflicted with a wide variety of variations that affect their performance. Under true operating conditions, the parameters chosen by the circuit designer are perturbed from their nominal values due to variations of various types. As a consequence, a single SPICE-level transistor or interconnect model, or its abstraction, is seldom an adequate predictor of the exact behavior of a circuit. These sources of variation can broadly be categorized into two classes. *Process variations* result from perturbations in the fabrication process, causing shifts from the nominal 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). *Environmental variations* arise from changes in the operating environment of the circuit, such as the temperature or variations in the supply voltage (V_{dd} and ground) levels, or the aging of the circuit. Both of these classes of variations can result in changes in the timing and power characteristics of a circuit.

Process variations can be classified into the following categories: die-to-die (D2D) variations are the variations from one die to another, while within-die (WID) variations correspond to variability within a single die. D2D variations affect all the devices on same chip in the same way, e.g., making the transistor gate lengths of devices on the same chip all larger or all smaller, while the WID variations may affect different devices differently on the same chip, e.g., making some devices have smaller transistor gate lengths and others larger transistor gate lengths. 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. In such a scenario, all variations of a specific parameter will cause the circuit delay to move in the same direction, and in the worst case, to a process corner. However, in the presence of WID variation, some parts of the circuit may speed up while others may slow down, and a more nuanced approach, based on statistical analysis, is necessary to capture these averaging effects.

The sources of these variations may be used to create another taxonomy: *Random variations* depict random behavior that can be characterized in terms of a distribution. This distribution 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 some 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

†1 University of Minnesota, Minneapolis, MN, USA

much spatial correlation at all, so that for all practical purposes, variations in neighboring transistors are uncorrelated. *Systematic variations* show predictable variational trends across a chip, and are caused by known physical phenomena during manufacturing. Examples of systematic variations include those due to (i) spatial WID gate length variability, also known as across-chip linewidth variation (ACLV), which observes systematic changes in the value of L_{eff} across a reticle due to effects such as changes in the stepper-induced illumination and imaging nonuniformity due to lens aberrations, 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.

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 here 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 great utility in increasing the design yield, and ultimately, the profit.

Process variations are "one-time variations" that affect the circuit at the time of manufacturing. A circuit that meets manufacturing test specifications is acceptable at the time of manufacturing, and its ability to meet these specifications due to process variations will not change over its lifetime. Simply put, parts that meet specifications are sent to the market, and those that do not can be discarded. As a result, statistical analysis is an excellent candidate for analyzing these variations, and can be used to drive economic considerations by averaging the cost of acceptable parts with those of the discarded parts. On the other hand, environmental variations are *run-time variations* that change over the lifetime of a part. For such shifts, it is more appropriate to use worst-case analyses that guarantee that the circuit functions correctly under any level of variation.

The remainder of this paper is organized as follows. We begin with a description of correlated variations on a die in Section 2. Next, we present an overview of methods for analyzing and optimizing the timing and power characteristics of a circuit under random process variations, in Section 3, followed by an account of environmental variations in Section 4.

2. Correlated Variations

The existence of correlations between WID variations complicates the task of statistical analysis. These correlations are of two types: spatial correlations, based on the spatial locality of objects, and structural correlations, which depend on the structure of the circuit.

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.

Spatial correlations among parameters model the fact that for some process parameters, devices or wires close to each other are more likely to have similar characteristics than those placed far away. The classical model for spatial correlation, which predicts the decay with distance, was proposed by Pelgrom [63]. 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 [3),11)] tessellate the die into n grid regions, as illustrated in **Fig. 1**. For a given process parameter, this assumes perfect correlations among for variations within a grid, high correlations for variations in nearby grids, and low or zero correlations in far-away grids. In the figure, gates a and b (whose sizes are shown to be exaggeratedly large) are located in the same grid square, and it is assumed that their parameter variations are always identical. Gates a and

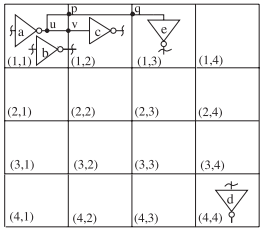

Fig. 1 A grid-based model for that captures spatial correlations on a die.

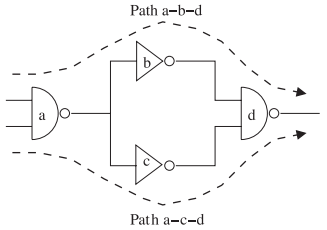

Fig. 2 An example to illustrate structural correlations in a circuit.

c lie in neighboring grids, and their parameter variations are not identical but highly correlated due to their spatial proximity (for example, when gate a has a larger than nominal gate length, it is highly probable that gate c will have a larger than nominal gate length, and less probable that it will have a smaller than nominal gate length). On the other hand, gates a and d are far away from each other, their parameters may be uncorrelated (e.g., when gate a has a larger than nominal gate length, the gate length for d may be either larger or smaller than nominal). Under this model, a parameter variation in a single grid at location (x, y) can be modeled using a single random variable $p(x, y)$, and the parameter distributions, including correlations, correspond to an n-variate distribution for each process variable with correlations.

For each type of parameter, n random variables are needed, each representing the value of a parameter in one of the n grids. Subsequent studies [48),96)] have investigated the development of techniques for characterizing these correlations, and a finer-grained model, based on the Karhunen-Loève expansion, is used for statistical circuit analysis in[10)].

Structural correlations arise from the structural properties of the circuit, and can be introduced through an example. Consider the reconvergent fanout structure shown in **Fig. 2**. The circuit has two paths, a-b-d and a-c-d. If, for example, we assume that each gate delay is a Gaussian random variable, then the probability density function (PDF) of the delay of each path is easy to compute, since it is the sum of Gaussians, which admits a closed form. However, the circuit delay is the maximum of the delays of these two paths, and these are correlated since the delays of a and d contribute to both paths. It is important to take such structural correlations, which arise due to reconvergences in the circuit, into account while performing timing and power analysis in the statistical context.

3. Random Process Variations

3.1 Statistical Static Timing Analysis

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 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 run-times. Therefore, more efficient methods are called for, based on SSTA.

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 Taylor series expansion as

$$D = D_0 + \sum_i \left[\frac{\partial f}{\partial p_i}\right]_0 \Delta p_i, \tag{1}$$

where D_0 is the nominal value of D, calculated at the nominal values of parameters in $\mathbf{p}$, the sensitivity $\left[\frac{\partial f}{\partial p_i}\right]_0$ 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. SSTA techniques can be classified in several ways:

- Path-based vs. block-based methods: Path-based methods are based on enumerating all paths in a circuit, unlike block based methods that perform a PERT-like topological traversal of the circuit. The latter class is significantly more computationally efficient, and is generally seen as the clear winner among the two.
- Continuous vs. discrete methods: This classification is based on whether the delay PDFs is assumed to be continuous or discrete.
- Gaussian vs. non-Gaussian modeling: This classification corresponds to the PDF used to represent the underlying variations. Some methods use closed form models, or assume knowledge of the moments of the variations, while others use interval-based analyses. Continuous methods often, but not always, use known PDF types (such as Gaussian PDFs) to model circuit delays. If the underlying parameters $p_i \in \mathbf{p}$ 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.
- Uncorrelated vs. correlated variations: If the delay PDFs are all assumed to be uncorrelated with each other, the task of SSTA is considerably simpler than when correlations are taken into account.

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 [74]. 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.

For the uncorrelated case, much of the analysis goes back to the work of Berkelaar, presented at the 1997 Tau workshop but never formally published until it was incorporated in a statistical gate sizing algorithm [37]. This work assumed that all gate delays are Gaussian, and presented a few key observations that are used in many of the SSTA methods that followed it. First, it observed that STA consists of the sum and max operations described above, and unlike the deterministic case, where the operands for both operations are fixed numbers, the operands in SSTA are PDFs, and must result in a PDF. Second, it was observed that the sum operator, operating on two Gaussian PDFs, produces a Gaussian result that is easy to characterize (the result of sum can be shown to be the convolution of the two PDFs), and the max operator applied to two Gaussians produces a non-Gaussian result. However, it is possible to compute the first two moments of the result of the max operation, and approximate the resulting PDF as a Gaussian with these moments. Results on benchmark circuits showed the viability of this method. This work was extended [89] to include intra-gate delay correlations: for this case, the result of the sum operator continues to remain a Gaussian, while the result of max was characterized using a classical technique [21]. Subsequent work for the uncorrelated case includes a method based on discrete PDFs [26], and methods that find tight upper and lower bounds for the accurate PDF [2),4),9)].

In the presence of correlated intradie variations, the analysis becomes significantly more complicated. Although approaches similar to early methods [89] can, in principle, be extended to correlated variations across gates, the associated computation requires every correlated variable to be compared with every other variable, resulting in prohibitive computation.

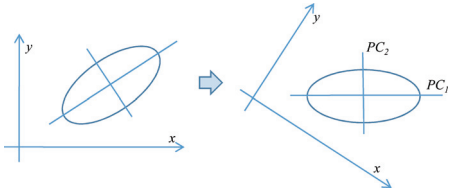

Fig. 3 An illustration of the idea of PCA: at left, a constant probability contour (provably an ellipse in 2D, or an ellipsoid in higher dimensions) of two jointly Gaussian random variables, x and y, is shown. A coordinate rotation can be employed to work with a new set of coordinates, PC_1 and PC_2, corresponding to the principal components. These are the axes of the ellipsoid.

A novel and simple method, based on the application of principal component analysis (PCA) techniques [56], was introduced [11),13)] to convert a set of correlated random variables into a set of uncorrelated variables in a transformed space, as illustrated in **Fig. 3**. The PCA step can be performed as a preprocessing step for a design, and depends purely on the technology and the gridding scheme used for modeling spatial correlation. The overall idea is similar to that of Berkelaar's, but the use of PCA overcomes the complexity bottleneck associated with using the correlated variables directly. In reality, some parameters may be spatially correlated and others (such as T_{ox} and N_d) may be uncorrelated: this method is easily extended to handle these issues. The complexity of the method is $p \cdot n$ times the complexity of an STA traversal of a circuit, where n is the number of squares in the grid and p is the number of correlated parameters, plus the complexity of finding the principal components, which requires very low runtimes in practice. A similar-looking method was presented subsequently [91].

SSTA methods can be classified on the basis of whether the Taylor series is truncated to a linear or a nonlinear approximation, and depending on whether the underlying variational parameters are assumed to be Gaussian or not. Of the four resulting classes, the linear Gaussian case has been addressed above. Several efforts have addressed extensions to more general cases. For the nonlinear Gaussian case a moment-based approach can be employed [46),47)]. 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 $p_i p_j$. A key property of this diagonalization is that it preserves the orthogonality of the principal components. This work was subsequently extended [102)] to develop a clever set of manipulations to compute the result of the max operator.

For the linear non-Gaussian case, an approach [79),80)] that transforms Gaussian parameter PDFs using PCA, and orthogonalizes non-Gaussian parameter PDFs using a procedure known as independent component analysis (ICA) [35)], 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 covers the most general case for performing statistical timing analysis. Several approaches [15),18),38),39)] 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 [46),47)] may be used and orthogonalized to remove cross terms of the type $p_i p_j$, the ICA transform that applies them to orthogonalized non-Gaussians [79),80)] can only guarantee that the PDFs in the transformed space will be uncorrelated, but not that they will be independent [⋆1]. 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.

A few noteworthy approaches to SSTA do not neatly fit into these categories. Interval-based analysis methods [50),51),81)] avoid using information about distributions, but propagate uncertainty intervals. These methods reduce the pessimism in interval propagation by transmitting not merely the interval, but also an affine functional approximation within the interval. However, they are more applicable to applications such as statistical interconnect analysis rather than SSTA: while they apply well to many arithmetic operators, they cannot easily handle

⋆1 Two random variables X and Y are uncorrelated if $E[XY] = E[X]E[Y]$, while they are independent if $E[f(X)g(Y)] = E[f(X)]E[g(Y)]$ for any functions f and g. For instance, if X and Y are independent, then $E[X^i Y^j] = E[X^i]E[Y^j]$. For Gaussian distributions, uncorrelatedness is identical to independence. For a general non-Gaussian distribution, independence implies uncorrelatedness, but not vice versa.

the max operator. A second class of approaches use statistically based methods to extend the corner-based paradigm to a statistical design scenario[57], or attempt pessimistic worst-case modeling [41),59)]. A third set of approaches use SSTA diagnostics to promote efficiency in post-silicon testing [49)].

3.2 Statistical 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 [44)]. Other work [70),71)] 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.

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 [61)]. 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 [70)].

In the presence of spatial variations, the central limit theorem does not hold since the variables fail to satisfy the requirement of independence that is necessary to apply the theorem. While the sum of lognormals is not a lognormal, it may be approximated as one using Wilkinson's method [1)]; 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 [12)] presents an approach for efficiently performing the addition using Wilkinson's method by reducing the effects of cross-terms. Another approach [83)] 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 [14)], and the resulting approach is shown to be better than either one individually.

3.3 Statistical Circuit 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.

The idea of statistical optimization is presented in **Fig. 4**, in a space where two design parameters, p_1 and p_2, may be varied. The upper picture shows the constant value contours of the objective function, and the feasible region where all constraints are met. The optimal value for the deterministic optimization problem is the point at which the lowest value contour intersects the feasible set, as shown. However, if there is a variation about this point that affects the objective function (assume, for simplicity here, that the constraints are unaffected by variations), then after manufacturing, the parameters may shift from the optimal design point. The figure shows an ellipsoidal variational region (corresponding to, say, the 99% probability contours of a Gaussian distribution) around an optimal design point: the manufactured solution may lie within this with a very high probability. It can be seen that a majority of points in this elliptical variational region lie outside the feasible set, implying a high likelihood that the manufactured circuit will fail the specifications. On the other hand, the robust optimum, shown in the lower picture, will ensure that the entire variational region will lie within the feasible set.

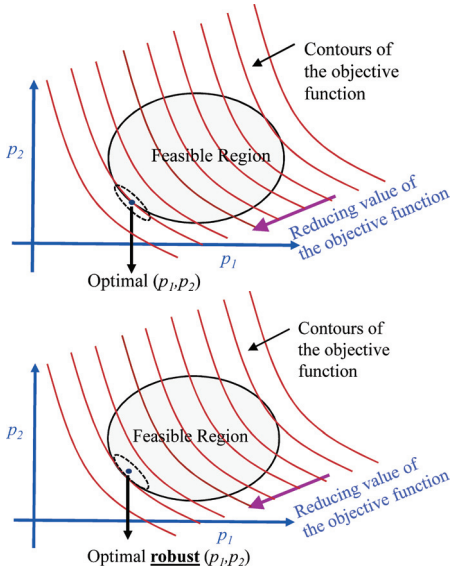

Fig. 4 A conceptual picture of robust optimization.

Next, we overview published research on the optimization problem of gate sizing in a statistical scenario. Early work [37], proposes formulation of statistical objective and timing constraints, and solves the resulting nonlinear optimization formulation. In other works on robust gate sizing [5),19),20),82)], 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 [69], 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 [52]. For the binning model, a yield optimization problem is formulated [25], 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 [76),80)]: 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),55),97)].

4. Environmental Variations

4.1 Temperature

Thermal problems are becoming increasingly important in affecting the behavior of digital circuits. This power dissipated on-chip generates heat that causes the on-chip temperatures to change, with some parts of the chip being hotter than others. Temperature and power (or heat flux) are intimately related, but it is important to note that they are distinct from each other.

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.

Conventional heat transfer on a chip is described by Fourier's law of conduc-

tion[⋆1], which states that the heat flux, q (in W/m^2), is proportional to the negative gradient of the temperature, T (in K), with the constant of proportionality corresponding to the thermal conductivity of the material, k_t (in W/(m K)). This leads to the thermal partial differential equation:

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

The boundary conditions for this equation are typically described in Dirichlet form, specifying information on the boundary of the chip.

The solution to Eq. (2) corresponds to the transient thermal response. In the steady state, all derivatives with respect to time go to zero, and therefore, steady-state analysis corresponds to solving the Poisson equation given by:

$$\nabla^2 T(\mathbf{r}) = -\frac{g(\mathbf{r})}{k_t} \quad (3)$$

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 thermal PDE can be solved using the finite difference method, where the thermal analysis can be shown to be equivalent to solving an RC circuit with current and voltage sources [60]. According to the thermal-electrical analogy, each node in the discretization corresponds to a node in the circuit. The steady-state 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.

⋆1 Nanoscale thermal analysis, which models electron-phonon interactions for fine-grained thermal analysis, involves the solution of the Boltzmann transport equation [65], but Fourier-based models are adequate for full-chip analysis.

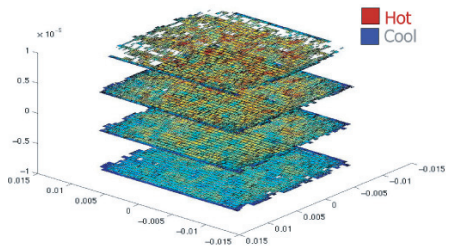

Fig. 5 A placement for the benchmark ibm01 in a four-tier 3D technology.

Other techniques for efficiently solving the thermal PDE for on-chip thermal analysis include techniques based on finite elements [28] and Green functions [92),101].

The on-chip thermal profile of a circuit can be optimized using a variety of techniques. These include altering the spatiotemporal distribution of power using microarchitectural optimization [32),33),58),73),95], and physical design techniques such as floorplanning [23),93),106], placement [17),22),28),30),88] and routing [6),24),103], and thermal via insertion [29]. Thermal mitigation to recover performance degradation due to thermal effects can be performed using adaptive body biases, adaptive supply voltages, and frequencies [8),27),53),54),64),98)–100].

An example of a thermally-driven optimization is thermally-driven 3D placement, where standard cells must be placed in a 3D chip with thermal constraints. The result of such an optimization is shown in **Fig. 5**, for the benchmark circuit, ibm01, in a four-tier 3D process. The cells are positioned in ordered rows on each tier, and the layout in each individual tier looks similar to a 2D standard cell layout. The heat sink is placed at the bottom of the 3D chip: the coolest cells are those in the bottom tier, next to the heat sink, and the temperature increases as we move to higher tiers.

4.2 Power Delivery

For correct circuit operation, it is essential to feed reliable values of the supply voltage, V_{dd}, and the ground voltage, typically 0V. Correct supply levels are essential to ensure that the logic value generated at gate outputs is correct. Moreover, a degraded V_{dd} leads to an increase in circuit delay. 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.

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)). DC solutions of these grids are useful in early stages of design, while transient solutions are necessary for more detailed analyses. Transient solutions may be computed either using time-stepping (constant time steps are typically used) or using model order reduction methods. For either DC or transient analysis, the set of equations to be solved correspond to a large system of linear equations, typically millions of variables. Time-domain techniques are popularly used in many tools, and several techniques for solving the analysis problem have been proposed. This system of equations is typically sparse and positive definite, but its large dimension necessitates the use of efficiency-enhancing methods. Chief among these are hierarchical methods, multigrid methods, and random walk based approaches.

Hierarchical methods [105] can 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. This step determines the voltages at the ports of the macromodels: each block is then solved using these voltage values as inputs. Since the sizes of the global grid (with the macromodels), as well as individual hierarchical blocks, are considerably smaller than the entire power grid, this method leads to large savings in computation time and memory usage.

The power grid problem looks similar to the finite difference discretization of a partial differential equation. Therefore, efficient methods from that domain may be used to solve the problem. Published approaches [40] employ the multigrid method to solve the problem. Multigrid methods 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. The procedure may be iterated to achieve further accuracy.

An analogy between random walks and power grids may be exploited to solve the network [66),67]. The key idea is that it is possible to obtain an estimate of the voltage at a node through a set of random walks starting at that node. Unlike most conventional techniques that require the entire network to be solved, even if the designer is only interested in the voltage at a single node, this technique provides a fast method for solving for the voltages in only a part of a network. Speedup techniques are incorporated, so that for medium levels of accuracy, the method can be faster than conventional direct methods. The idea is extended [68] to create a preconditioner for a direct solver, and a theoretical link between the random walk method and LU factorization is presented.

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 [75),104], topology optimization [77),78], wire sizing [85),86], and decoupling capacitor (decap) insertion [87]. 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.

4.3 Aging and Reliability

During the lifetime of a circuit, stresses caused during normal operation can cause its performance to degrade. We will outline several significant reliability effects here, namely, bias temperature instability and oxide breakdown.

4.3.1 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 phenomenon of negative bias temperature instability (NBTI) can be illustrated with the help of a simple circuit, an inverter, illustrated in **Fig. 6** (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

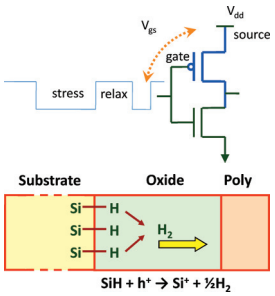

Fig. 6 An inverter whose PMOS device is alternately subjected to NBTI stress and relax phases, and an illustration of the phenomenon of NBTI.

in Fig. 6 (b), causing the threshold voltage to degrade, slowing down the gate, and potentially causing the circuit to fail to meet specifications. When the stress is removed, the number of interface traps is reduced, and the threshold voltage recovers towards its original value. Several models for NBTI have been proposed [7),31),34),36),42),90)].

A corresponding and dual effect, known as Positive Bias Temperature Instability (PBTI) can be seen for NMOS devices. Although PBTI causes less degradation than NBTI [72)], it is becoming increasingly important in its own right.

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. An analytical model for the change in threshold voltage over multiple stress-relax cycles is provided [42)].

The degradation in threshold voltage shows a property known as frequency-independence, demonstrated over a wide range of frequencies [7),16),42)]: if a pattern of signals is applied to a transistor over time, the degradation depends only on the total fraction of time for which the transistor was stressed, and not on the frequency of the signal, or the distribution of stress/relax times. Accordingly, if one defines a signal probability of the signal value at the gate node of the transistor, corresponding to the proportion of time that the transistor is likely to be under stress, the threshold voltage degradation is only a function of this signal probability, and a one-dimensional look-up table can be used to store this degradation. Techniques for guard-banding a circuit for NBTI degradations have been proposed [43),62),90)].

4.3.2 Oxide Breakdown

Time-dependent dielectric breakdown (TDDB) 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 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 [84),94)]. 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

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

The parameter α corresponds to the time-to-breakdown at about the 63rd percentile, and β is the Weibull slope. Generally speaking, an increased electric field (i.e., an increased voltage across the gate) accelerates breakdown. Currently, there are few approaches to addressing oxide breakdown issues at the circuit level, and this is a significant open problem for CAD researchers.

5. Conclusion

This paper has presented an overview of the challenges and issues that arise due to on-chip variability, the corresponding need for statistical design techniques, and a summary of the current state of research in this area. The description presented here is merely a beginning, and this is an area of intense activity and

research. The reader is invited to explore deeper, and the extensive reference list is intended to assist with this.

References

1) Abu-Dayya, A.A. and Beaulieu, N.C.: Comparison of methods of computing correlated lognormal sum distributions and outages for digital wireless applications, *Proc. IEEE Vehicular Technology Conference*, pp.175–179 (1994).
2) Agarwal, A., Blaauw, D. and Zolotov, V.: Statistical timing analysis for intradie process variations with spatial correlations, *Proc. IEEE/ACM International Conference on Computer-Aided Design*, pp.900–907 (2003).
3) Agarwal, A., Blaauw, D., Zolotov, V., Sundareswaran, S., Zhao, M., Gala, K. and Panda, R.: Statistical delay computation considering spatial correlations, *Proc. Asia-South Pacific Design Automation Conference*, Kitakyushu, Japan, pp.271–276 (2003).
4) Agarwal, A., Blaauw, D., Zolotov, V. and Vrudhula, S.: Computation and refinement of statistical bounds on circuit delay, *Proc. ACM/IEEE Design Automation Conference*, pp.348–353 (2003).
5) Agarwal, A., Chopra, K., Blaauw, D. and Zolotov, V.: Circuit optimization using statistical static timing analysis, *Proc. ACM/IEEE Design Automation Conference*, pp.338–342 (2005).
6) Ajami, A.H., Banerjee, K. and Pedram, M.: Modeling and analysis of nonuniform substrate temperature effects on global ULSI interconnects, *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, Vol.24, No.6, pp.849–861 (2005).
7) Alam, M.A.: A critical examination of the mechanics of dynamic NBTI for pMOSFETs, *IEEE International Electronic Devices Meeting*, pp.14.4.1–14.4.4 (2003).
8) Andrei, A., Schmitz, M., Eles, P., Peng, Z. and Al-Hashimi, B.M.: Overhead-conscious voltage selection for dynamic and leakage energy reduction of time-constrained systems, *Proc. Design, Automation & Test in Europe*, pp.518–523 (2004).
9) Bhardwaj, S., Vrudhula, S. and Blaauw, D.: TAU: Timing analysis under uncertainty, *Proc. IEEE/ACM International Conference on Computer-Aided Design*, pp.615–620 (2003).
10) Bhardwaj, S., Vrudhula, S., Ghanta, P. and Cao, Y.: Modeling of intradie process variations for accurate analysis and optimization of nanoscale circuits, *Proc. ACM/IEEE Design Automation Conference*, pp.791–796 (2006).
11) Chang, H. and Sapatnekar, S.S.: Statistical timing analysis considering spatial correlations using a single PERT-like traversal, *Proc. IEEE/ACM International Conference on Computer-Aided Design*, pp.621–625 (2003).
12) Chang, H. and Sapatnekar, S.S.: Full-chip analysis of leakage power under process variations, including spatial correlations, *Proc. ACM/IEEE Design Automation Conference*, pp.523–528 (2005).
13) Chang, H. and Sapatnekar, S.S.: Statistical timing analysis under spatial correlations, *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, Vol.24, No.9, pp.1467–1482 (2005).
14) Chang, H. and Sapatnekar, S.S.: Prediction of leakage power under process uncertainties, *Proc. ACM Transactions on Design Automation of Electronic Systems*, Vol.12, No.2 (2007). Article 12 (27 pages).
15) Chang, H., Zolotov, V., Narayan, S. and Visweswariah, C.: Parameterized block-based statistical timing analysis with non-Gaussian parameters, nonlinear delay functions, *Proc. ACM/IEEE Design Automation Conference*, pp.71–76 (2005).
16) Chen, G., Chuah, K.Y., Li, M.F., Chan, D.S.H., Ang, C.H., Cheng, J.Z., Jin, Y. and Kwong, D.L.: Dynamic NBTI of PMOS transistors and its impact on device lifetime, *Proc. IEEE International Reliability Physics Symposium*, pp.196–200 (2003).
17) Chen, G. and Sapatnekar, S.S.: Partition-driven standard cell placement, *Proc. International Symposium on Physical Design*, pp.75–80 (2003).
18) Cheng, L., Xiong, J. and He, L.: Non-linear statistical static timing analysis for non-gaussian variation sources, *Proc. ACM/IEEE Design Automation Conference*, pp.250–255 (2007).
19) Choi, S.H., Paul, B.C. and Roy, K.: Novel sizing algorithm for yield improvement under process variation in nanometer technology, *Proc. ACM/IEEE Design Automation Conference*, pp.454–459 (2004).
20) Chopra, K., Shah, S., Srivastava, A., Blaauw, D. and Sylvester, D.: Parametric yield maximization using gate sizing based on efficient statistical power and delay gradient computation, *Proc. IEEE/ACM International Conference on Computer-Aided Design*, pp.1023–1028 (2005).
21) Clark, C.: The greatest of a finite set of random variables, *Operations Research*, Vol.9, pp.85–91 (1961).
22) Cong, J., Luo, G., Wei, J. and Zhang, Y.: Thermal-aware 3D IC placement via transformation, *Proc. Asia-South Pacific Design Automation Conference*, pp.780–785 (2007).
23) Cong, J., Wei, J. and Zhang, Y.: A thermal-driven floorplanning algorithm for 3D ICs, *Proc. International Symposium on Physical Design*, pp.306–313 (2004).
24) Cong, J. and Zhang, Y.: Thermal-driven multilevel routing for 3-D ICs, *Proc. Asia-South Pacific Design Automation Conference*, pp.121–126 (2005).
25) Davoodi, A. and Srivastava, A.: Variability driven gate sizing for binning yield optimization, *Proc. ACM/IEEE Design Automation Conference*, pp.959–964 (2006).
26) Devgan, A. and Kashyap, C.: Block-based static timing analysis with uncertainty, *Proc. IEEE/ACM International Conference on Computer-Aided Design*, pp.607–614 (2003).

27) Fischer, T., Anderson, F., Patella, B. and Naffziger, S.: A 90nm variable-frequency clock system for a power-managed Itanium®-family processor, *Proc. IEEE International Solid-State Circuits Conference*, pp.294–299,599 (2005).
28) Goplen, B. and Sapatnekar, S.S.: Efficient thermal placement of standard cells in 3D ICs using a force directed approach, *Proc. IEEE/ACM International Conference on Computer-Aided Design*, pp.86–89 (2003).
29) Goplen, B. and Sapatnekar, S.S.: Thermal via placement in 3D ICs, *Proc. International Symposium on Physical Design*, pp.167–174 (2005).
30) Goplen, B. and Sapatnekar, S.S.: Placement of 3D ICs with thermal and interlayer via considerations, *Proc. ACM/IEEE Design Automation Conference*, pp.626–631 (2007).
31) Grasser, T., Gos, W., Sverdlov, V. and Kaczer, B.: The universality of NBTI relaxation and its implications for modeling and characterization, *Proc. IEEE International Reliability Physics Symposium*, pp.268–280 (2007).
32) Han, Y., Koren, I. and Moritz, C.A.: Temperature aware floorplanning, *Second Workshop on Temperature-Aware Computing Systems* (2005).
33) Healy, M., Vittes, M., Ekpanyapong, M., Ballapuram, C., Lim, S.K., Lee, H.-H.S. and Loh, G.H.: Microarchitectural floorplanning under performance and thermal tradeoff, *Proc. Design, Automation & Test in Europe*, pp.1–6 (2006).
34) Huard, V., Denais, M. and Parthasarathy, C.: NBTI degradation: From physical mechanisms to modeling, *Journal of Microelectronics Reliability*, Vol.46, pp.1–23 (2006).
35) Hyvärinen, A. and Oja, E.: Independent component analysis: Algorithms and applications, *Neural Networks*, Vol.13, pp.411–430 (2000).
36) Islam, A.E., Kufluoglu, H., Varghese, D., Mahapatra, S. and Alam, M.A.: Recent issues in negative bias temperature instability: Initial degradation, field dependence of interface trap generation, hole trapping effects, and relaxation, *IEEE Transactions on Electron Devices*, Vol.54, pp.2143–2154 (2007).
37) Jacobs, E. and Berkelaar, M.: Gate sizing using a statistical delay model, *Proc. Design, Automation & Test in Europe*, pp.283–290 (2000).
38) Khandelwal, V. and Srivastava, A.: A general framework for accurate statistical timing analysis considering correlations, *Proc. ACM/IEEE Design Automation Conference*, pp.89–94 (2005).
39) Khandelwal, V. and Srivastava, A.: A quadratic modeling-based framework for accurate statistical timing analysis considering correlations, *IEEE Transactions on VLSI Systems*, Vol.15, No.2, pp.206–215 (2007).
40) Kozhaya, J., Nassif, S.R. and Najm, F.N.: A multigrid-like technique for power grid analysis, *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, Vol.21, No.10, pp.1148–1160 (2002).
41) Kumar, S.V., Kashyap, C.V. and Sapatnekar, S.S.: A framework for block-based timing sensitivity analysis, *Proc. ACM/IEEE Design Automation Conference* (2008).
42) Kumar, S.V., Kim, C.H., and Sapatnekar, S.S.: An analytical model for negative bias temperature instability (NBTI), *Proc. IEEE/ACM International Conference on Computer-Aided Design*, pp.493–496 (2006).
43) Kumar, S.V., Kim, C.H. and Sapatnekar, S.S.: NBTI-aware synthesis of digital circuits, *Proc. ACM/IEEE Design Automation Conference*, pp.370–375 (2007).
44) Lee, D., Kwong, W., Blaauw, D. and Sylvester, D.: Analysis and minimization techniques for total leakage considering gate oxide leakage, *Proc. ACM/IEEE Design Automation Conference*, pp.175–180 (2003).
45) Li, X., Le, J., Celik, M. and Pileggi, L.T.: Defining statistical sensitivity for timing optimization of logic circuits with large-scale process and environmental variations, *Proc. IEEE/ACM International Conference on Computer-Aided Design*, pp.844–851 (2005).
46) Li, X., Le, J., Gopalakrishnan, P. and Pileggi, L.T.: Asymptotic probability extraction for non-normal distributions of circuit performance, *Proc. IEEE/ACM International Conference on Computer-Aided Design*, pp.2–9 (2004).
47) Li, X., Le, J., Gopalakrishnan, P. and Pileggi, L.T.: Asymptotic probability extraction for nonnormal performance distributions, *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, Vol.26, No.1, pp.16–37 (2007).
48) Liu, F.: A general framework for spatial correlation modeling in VLSI design, *Proc. ACM/IEEE Design Automation Conference*, pp.817–822 (2007).
49) Liu, Q. and Sapatnekar, S.S.: Confidence scalable post-silicon statistical delay prediction under process variations, *Proc. ACM/IEEE Design Automation Conference*, pp.497–502 (2007).
50) Ma, J.D. and Rutenbar, R.A.: Fast interval-valued statistical modeling of interconnect and effective capacitance, *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, Vol.25, No.4, pp.710–724 (2006).
51) Ma, J.D. and Rutenbar, R.A.: Interval-valued reduced order statistical interconnect modeling, *Proc. IEEE/ACM International Conference on Computer-Aided Design*, pp.460–467 (2006).
52) Mani, M., Devgan, A. and Orshansky, M.: An efficient algorithm for statistical power under timing yield constraints, *Proc. ACM/IEEE Design Automation Conference*, pp.309–314 (2005).
53) Martin, S.M., Flautner, K., Mudge, T. and Blaauw, D.: Combined dynamic voltage scaling and adaptive body biasing for lower power microprocessors under dynamic workloads, *Proc. IEEE/ACM International Conference on Computer-Aided Design*, pp.721–725 (2002).
54) McGowen, R., Poirier, C.A., Bostak, C., Ignowski, J., Millican, M., Parks, W.H. and Naffziger, S.: Power and temperature control on a 90-nm Itanium family processor, *IEEE Journal of Solid-State Circuits*, Vol.41, No.1, pp.229–237 (2006).
55) Mogal, H., Qian, H., Sapatnekar, S.S. and Bazargan, K.: Clustering based pruning

for statistical criticality computation under process variations, *Proc. IEEE/ACM International Conference on Computer-Aided Design*, pp.340–343 (2007).
56) Morrison, D.: *Multivariate Statistical Methods*, McGraw-Hill, New York, NY (1976).
57) Najm, F.N., Menezes, N. and Ferzli, I.A.: 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, No.3, pp.574–591 (2007).
58) Nookala, V., Lilja, D.J. and Sapatnekar, S.S.: Temperature-aware floorplanning of microarchitecture blocks with IPC-power dependence modeling and transient analysis, *Proc. ACM International Symposium on Low Power Electronics and Design*, pp.298–303 (2006).
59) Onaissi, S. and Najm, F.N.: A linear-time approach for static timing analysis covering all process corners, *Proc. IEEE/ACM International Conference on Computer-Aided Design*, pp.217–224 (2006).
60) Ozisik, M.N.: *Finite Difference Methods in Heat Transfer*, CRC Press, New York, New York, USA (1994).
61) Papoulis, A. and Pillai, S.U.: *Probability, Random Variables, and Stochastic Processes*, McGraw-Hill, Boston, Massachusetts, USA (2002).
62) Paul, B.C., Kang, K., Kufluoglu, H., Alam, M.A. and Roy, K.: Temporal performance degradation under NBTI: Estimation and design for improved reliability of nanoscale circuits, *Proc. Design, Automation & Test in Europe*, pp.1–6 (2006).
63) Pelgrom, M.J.M., Duinmaijer, A.C.J. and Welbers, A.P.G.: Matching properties of MOS transistors, *IEEE Journal of Solid-State Circuits*, Vol.24, No.5, pp.1433–1439 (1989).
64) Piorier, C., McGowen, R., Bostak, C. and Naffziger, S.: Power and temperature control on an Itanium®-family processor, *Proc. IEEE International Solid-State Circuits Conference*, pp.304–305 (2005).
65) Pop, E., Sinha, S. and Goodson, K.E.: Heat generation and transport in nanometer-scale transistors, *Proc. IEEE*, Vol.94, No.8, pp.1587–1601 (2006).
66) Qian, H., Nassif, S.R. and Sapatnekar, S.S.: Random walks in a supply network, *Proc. ACM/IEEE Design Automation Conference*, pp.93–98 (2003).
67) Qian, H., Nassif, S.R. and Sapatnekar, S.S.: Power grid analysis using random walks, *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, Vol.24, No.8, pp.1204–1224 (2005).
68) Qian, H. and Sapatnekar, S.S.: A hybrid linear equation solver and its application in quadratic placement, *Proc. IEEE/ACM International Conference on Computer-Aided Design*, pp.905–909 (2005).
69) Raj, S., Vrudhala, S.B.K. and Wang, J.: A Methodology to Improve Timing Yield in the Presence of Process Variations, *Proc. ACM/IEEE Design Automation Conference*, pp.448–453 (2004).
70) Rao, R., Devgan, A., Blaauw, D. and Sylvester, D.: Parametric yield estimation considering leakage variability, *Proc. ACM/IEEE Design Automation Conference*, pp.442–447 (2004).
71) Rao, R., Srivastava, A., Blaauw, D. and Sylvester, D.: Statistical estimation of leakage current considering inter- and intra-die process variation, *Proc. ACM International Symposium on Low Power Electronics and Design*, pp.84–89 (2003).
72) Reddy, V., Krishnan, A.T., Marshall, A., Rodriguez, J., Natarajan, S., Rost, T. and Krishnan, S.: Impact of negative bias temperature instability on digital circuit reliability, *Proc. IEEE International Reliability Physics Symposium*, pp.248–254 (2002).
73) Sankaranarayanan, K., Velusamy, S., Stan, M. and Skadron, K.: A case for thermal-aware floorplanning at the microarchitectural level, *The Journal of Instruction-Level Parallelism*, Vol.8 (2005).
74) Sapatnekar, S.S.: *Timing*, Springer, Boston, Massachusetts, USA (2004).
75) Sato, T., Onodera, H. and Hashimoto, M.: Successive pad assignment algorithm to optimize number and location of power supply pad using incremental matrix inversion, *Proc. Asia-South Pacific Design Automation Conference*, pp.723–728 (2005).
76) Singh, J., Nookala, V., Luo, T. and Sapatnekar, S.: Robust gate sizing by geometric programming, *Proc. ACM/IEEE Design Automation Conference*, pp.315–320 (2005).
77) Singh, J. and Sapatnekar, S.S.: Congestion-aware topology optimization of structured power/ground networks, *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, Vol.24, No.5, pp.683–695 (2005).
78) Singh, J. and Sapatnekar, S.S.: A partition-based algorithm for power grid design using locality, *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, Vol.25, No.4, pp.664–677 (2006).
79) Singh, J. and Sapatnekar, S.S.: Statistical timing analysis with correlated non-Gaussian parameters using independent component analysis, *Proc. ACM/IEEE Design Automation Conference*, pp.155–160 (2006).
80) Singh, J. and Sapatnekar, S.S.: 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, No.1, pp.160–173 (2008).
81) Singhee, A., Fang, C.R., Ma, J.D. and Rutenbar, R.A.: Probabilistic interval-valued computation: toward a practical surrogate for statistics inside CAD tools, *Proc. ACM/IEEE Design Automation Conference*, pp.167–172 (2006).
82) Sinha, D., Shenoy, N.V. and Zhou, H.: Statistical gate sizing for timing yield optimization, *Proc. IEEE/ACM International Conference on Computer-Aided Design*, pp.1037–1042 (2005).
83) Srivastava, A., Shah, S., Agarwal, K., Sylvester, D., Blaauw, D. and Director, S.W.: Accurate and efficient gate-level parametric yield estimation considering correlated variations in leakage power and performance, *Proc. ACM/IEEE Design Automation*

Conference, pp.535–540 (2005).
84) Stathis, J.H.: Reliability limites for the gate insulator in CMOS technology, *IBM Journal of Research and Development*, Vol.46, No.2/3, pp.265–286 (2002).
85) Su, H., Hu, J., Nassif, S.R. and Sapatnekar, S.S.: Congestion-driven codesign of power and signal networks, *Proc. ACM/IEEE Design Automation Conference*, pp.477–480 (2002).
86) Su, H., Hu, J., Sapatnekar, S.S. and Nassif, S.R.: A methodology for the simultaneous design of supply and signal networks, *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, Vol.23, No.12, pp.1614–1624 (2004).
87) Su, H., Sapatnekar, S.S. and Nassif, S.R.: Optimal decoupling capacitor sizing and placement for standard cell layout designs, *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, Vol.22, No.4, pp.428–436 (2003).
88) Tsai, C.H. and Kang, S.M.: Cell-level placement for improving substrate thermal distribution, *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, Vol.19, No.2, pp.253–266 (2000).
89) Tsukiyama, S., Tanaka, M. and Fukui, M.: A statistical static timing analysis considering correlations between delays, *Proc. Asia-South Pacific Design Automation Conference*, pp.353–358 (2001).
90) Vattikonda, R., Wang, W. and Cao, Y.: Modeling and minimization of PMOS NBTI effect for robust nanometer design, *Proc. ACM/IEEE Design Automation Conference*, pp.1047–1052 (2006).
91) Visweswariah, C., Ravindran, K., Kalafala, K., Walker, S.G. and Narayan, S.: First-order incremental block-based statistical timing analysis, *Proc. ACM/IEEE Design Automation Conference*, pp.331–336 (2004).
92) Wang, B. and Mazumder, P.: Accelerated chip-level thermal analysis using multilayer Green's function, *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, Vol.26, No.2, pp.325–344 (2007).
93) Wong, E. and Lim, S.K.: 3D floorplanning with thermal vias, *Proc. Design, Automation & Test in Europe*, pp.878–883 (2006).
94) Wu, E.Y., Nowak, E.J., Vayshenker, A., Lai, W.L. and Harmon, D.L.: CMOS scaling beyond the 100-nm node with silicon-dioxide-based gate dielectrics, *IBM Journal of Research and Development*, Vol.46, No.2/3, pp.287–298 (2002).
95) Wu, Y.W., Yang, C.-L., Yuh, P.-H. and Chang, Y.-W.: Joint exploration of architectural and physical design spaces with thermal consideration, *Proc. ACM International Symposium on Low Power Electronics and Design*, pp.123–126 (2005).
96) Xiong, J., Zolotov, V. and He, L.: Robust extraction of spatial correlation, *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, Vol.26, No.4, pp.619–631 (2007).
97) Xiong, J., Zolotov, V., Venkateswaran, N. and Visweswariah, C.: Criticality computation in parameterized statistical timing, *Proc. ACM/IEEE Design Automation Conference*, pp.63–68 (2006).
98) Yajuan, S., Zuodong, W. and Shaojun, W.: Energy-aware supply and body biasing voltage scheduling algorithm, *Proc. International Conference on Solid State and Integrated Circuits Technology*, pp.1956–1959 (2004).
99) Yan, L., Luo, J. and Jha, N.K.: Combined dynamic voltage scaling and adaptive body biasing for heterogeneous distributed real-time embedded systems, *Proc. IEEE/ACM International Conference on Computer-Aided Design*, pp.30–37 (2003).
100) Yan, L., Luo, J. and Jha, N.K.: 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, No.7, pp.1030–1041 (2005).
101) Zhan, Y. and Sapatnekar, S.S.: High efficiency Green function-based thermal simulation algorithms, *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, Vol.26, No.9, pp.1661–1675 (2007).
102) Zhan, Y., Strojwas, A.J., Li, X., Pileggi, L.T., Newmark, D. and Sharma, M.: Correlation-aware statistical timing analysis with non-Gaussian delay distributions, *Proc. ACM/IEEE Design Automation Conference*, pp.77–82 (2005).
103) Zhang, T., Zhan, Y. and Sapatnekar, S.S.: Temperature-aware routing in 3D ICs, *Proc. Asia-South Pacific Design Automation Conference*, pp.309–314 (2006).
104) Zhao, M., Fu, Y., Zolotov, V., Sundareswaran, S. and Panda, R.: Optimal placement of power-supply pads and pins, *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, Vol.25, No.1, pp.144–154 (2006).
105) Zhao, M., Panda, R.V., Sapatnekar, S.S. and Blaauw, D.: Hierarchical analysis of power distribution networks, *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, Vol.21, No.2, pp.159–168 (2002).
106) Zhou, P., Ma, Y., Li, Z., Dick, R.P., Shang, L., Zhou, H., Hong, X. and Zhou, Q.: 3D-STAF: Scalable temperature and leakage aware floorplanning for three-dimensional integrated circuits, *Proc. IEEE/ACM International Conference on Computer-Aided Design*, pp.590–597 (2007).

(Received March 24, 2008)
(Released August 27, 2008)

(Invited by Editor-in-Chief: *Hidetoshi Onodera*)

Sachin Sapatnekar received the Ph.D. degree from the University of Illinois at Urbana-Champaign in 1992. He is currently a Professor in the Department of Electrical and Computer Engineering at the University of Minnesota, where he holds the Robert and Marjorie Henle Chair and the Distinguished McKnight University Professorship. He has authored several books and papers in the areas of timing and layout, and has held positions on the editorial board of the *IEEE Transactions on CAD*, the *IEEE Transactions on Circuits and Systems II*, *IEEE Design & Test*, and the *IEEE Transactions on VLSI Systems*. He has served on the Technical Program Committee for various conferences, and as Technical Program and General Chair for the *ACM International Symposium on Physical Design* and the *IEEE Workshop on Timing Issues in the Specification and Synthesis in Digital Systems (Tau)*, and Techical Program Co-chair for the *ACM/IEEE Design Automation Conference*. He is a recipient of the NSF Career Award, three best paper awards at DAC and one at ICCD, and the SRC Technical Excellence award, and is a fellow of the IEEE.