Power Dissipation, Variations and Nanoscale CMOS Design: Test Challenges and Self-Calibration/Self-Repair Solutions

Swarup Bhunia[*] and Kaushik Roy[$]
[*]Dept of EECS, Case Western Reserve University, [$]Dept of ECE, Purdue University
skb21@case.edu, kaushik@ecn.purdue.edu

Abstract

In the nanometer technology regime, power dissipation and process parameter variations have emerged as major design considerations. These problems continue to grow with leakage power becoming a dominant form of power consumption. On the other hand, variations in the device parameters, both systematic and random, translate into variations in circuit parameters like delay and leakage, leading to loss in parametric yield. Numerous design techniques have been investigated for both logic and memory circuits to address the growing issues with power and variations. Low-power and process-tolerant designs, however, impose new test challenges and may even have conflicting requirements for test – affecting delay fault coverage, I_{DDQ} testability, parametric yield, and even stuck-at tests. Hence, there is a need to consider test and yield, while designing for low-power and robustness under variations. In this paper, we provide an overview of major low-power and variation-tolerant design techniques; discuss related test issues and focus on effectiveness of self-calibration/self-repair solutions to maintain high yield while achieving low power dissipation.

1. Introduction

At nanometer-scale geometry, power dissipation and process parameter variations have emerged as major barriers to gigascale integration [1-2]. Although dynamic power traditionally has been the significant form of power consumption in sub-micron process nodes, aggressive technology scaling has exposed the secondary problem of leakage power [3], which contributes to nearly 20-40% of total power in deep sub-micron modern microprocessors [36]. Increased power dissipation also manifests as increase in junction temperature due to limited cooling capacity of the package. To improve battery-life in portable devices and to reduce temperature-induced reliability concerns, numerous power saving techniques have been investigated at circuit and architecture level that target reduction of leakage and/or dynamic power. Due to quadratic dependence of dynamic power on supply voltage, voltage scaling has emerged as a popular choice for dynamic power reduction. Besides scaling of supply voltage, other important low-power design techniques that target dynamic power reduction are: gate sizing for reduction in effective switching capacitance, clock gating, and frequency scaling. On the other hand, dominant leakage saving techniques for logic and memory circuits include transistor stacking, dual or multiple threshold voltage CMOS and body biasing. Although these techniques provide effective power saving solutions, many of these techniques cause undesirable consequences on test and parametric yield of the design.

Another major design challenge in the nanometer regime is increased process parameter variations [2, 4-14]. Process imperfections due to sub-wavelength lithography lead to device level variations in small-geometry devices. Variations in device parameters such as length, width, oxide thickness, and flat-band voltage of devices along with random dopant fluctuations (RDF) and line edge roughness (LER) are making the devices exhibit large variations in their circuit parameters, particularly in the threshold voltage (V_{th}). Threshold voltage is a strong determinant of circuit speed: low-V_{th} chips are typically faster than high-V_{th} ones (since low-V_{th} corresponds to higher drive current). Statistical variations in device parameters lead to a statistical distribution of V_{th}. Consequently, delay of a circuit (and thus the maximum allowable frequency of operation) also follows a statistical distribution [6, 8]. Hence, parametric yield of a circuit (probability to meet the desired performance or power specification) is expected to suffer considerably, unless an overly pessimistic worst-case design approach is followed. Since leakage power of a circuit has exponential dependence on device threshold voltage (V_{th}), parameter variations results in large variability in leakage power [12, 18] along with variation in circuit delay. Moreover, threshold voltage variation poses concern in robustness of operation, particularly in Static Random Access Memory (SRAM) and dynamic logic circuits (such as domino).

Since worst-case design approach may incur prohibitive design overhead, multitude of research efforts have been devoted to explore alternative design methodologies under variations. Broadly, three classes of techniques are proposed to ensure/enhance yield under variations while incurring minimal design overhead: 1) *Statistical design approach,* where a circuit parameter (e.g. delay or leakage) is modeled as a statistical distribution (e.g. Gaussian) and the circuit is designed to meet a constraint on yield (or to maximize it) with respect to a target value of the parameter [4-5, 9, 12]. Gate sizing or dual-V_{th} CMOS are examples of techniques that can be used to vary circuit delay or leakage distribution. 2) *Variation avoidance*, where a given circuit is synthesized using nominal parameter values, however, any possible failures due to delay variations are identified at run time and avoided by adaptively switching to two-cycle operations [19]. 3) *Post-Silicon compensation and correction*, where parameter shift is detected (using delay or leakage sensor) and adjusted after manufacturing by changing operating

1-4244-1128-9/07/$25.00 © 2007 IEEE

parameters such as supply voltage, frequency or body bias.

Variations in process parameters (in particular, threshold voltage) can also lead to failures in a Static Random Access Memory (SRAM) array, degrading memory yield [13-14]. Intra-die process variation is a major concern for memory design since it introduces mismatch in strength between two identical transistors in a memory cell. Similar to logic circuit, different circuit and architecture level design techniques have been investigated [16, 20] to improve yield of nanoscaled SRAM.

Parameter variations can have large negative impact on test affecting both test-quality and cost [24-26]. In particular, delay testing under probabilistic path delay model can be challenging in terms of path selection and pattern generation for path sensitization [30]. Parameter variations also affect noise margin of dynamic circuits, which in turn puts burden on test to check robustness of these circuits after manufacturing. The combined impact of advanced power management techniques (such as dynamic voltage scaling or clock gating) and process-induced uncertainty in device parameters bring new challenges to conventional ATE-based testing. One of the difficulties is to mimic the worst-case operating condition during test. Considering the large number of operating points in today's high-performance chips (defined by supply voltage, frequency and temperature), ensuring correct operation under all possible conditions has become a major test challenge. Low-power and process-tolerant design techniques may also have conflicting requirements for test. Hence, there is a need to consider test and yield, while designing for low-power and variation tolerance.

In this paper, we highlight the major test challenges associated with nanoscale CMOS designs. In particular, we discuss test challenges related to low-power and variation-tolerant designs and focus on a new class of design techniques based on self-calibration and self-repair that can potentially reduce burden on test and help achieve increased test confidence and higher yield.

The rest of the article is organized as follows. Section 2 presents major techniques for low-power logic and memory design and associated test considerations. In Section 3, we discuss robust design under process variations with their test impact. In Section 4, we analyze design techniques for improving yield and reliability under variations using self-calibration/self-repair. Section 5 concludes the article.

2. Power-Conscious Design and Test

2.1 Leakage Power

Increasing leakage power with technology scaling poses both design and test concerns. The leakage current in a nanoscale transistor has several components [3], which are shown in Fig. 1. Transistor off-state current (I_{OFF}) is the drain current when the gate-to-source voltage is 0. The components that influence I_{OFF} include the threshold voltage, the channel's physical dimensions, the channel and surface doping profiles, the drain-source junction depth, the gate-oxide thickness, and V_{DD}. In long-channel devices, leakage from drain-substrate, source-substrate, and well-substrate reverse-bias p-n junctions (I_1) dominates I_{OFF} and is negligible. However, in scaled devices, the halo implants in the source and drain junctions can lead to large band-to-band junction tunneling current. The subthreshold current (I_2) is another dominant component of leakage and is even important in the active mode of operation due to its strong dependence on temperature. For short channel devices, drain-induced barrier lowering (DIBL) modulates I_2. Because of thin gate-oxide, the tunneling current through the oxide (I_3), referred as *gate leakage*, can be large in scaled technologies. The current due to hot carrier effect (I_4), the punch-through current (I_5), and the gate-induced drain leakage (I_6) can all significantly affect total leakage.

The total contribution of all leakage components constitutes a major source of power dissipation in sub-100nm logic and memory circuits. While increasing leakage power has triggered circuit and architecture level leakage control techniques, it has also affected design testability significantly. Two major impacts of increasing leakage on testability are: 1) ***I_{DDQ}* Testability:** Technology scaling challenges the effectiveness of current-based test techniques such as I_{DDQ} testing. Sensitivity of I_{DDQ} testing reduces drastically due to high intrinsic leakage. 2) **Impact on Burn-In:** The exponential dependence of subthreshold leakage on temperature leads to positive feedback that can result in thermal runaway condition and yield loss during burn-in test (when stressed voltage and temperature are applied).

2.2 Leakage Control Mechanisms

Leakage control techniques (such as input vector control, supply gating, and multiple-threshold design) can have positive affect on I_{DDQ} test and burn-in. Next, we discuss some major leakage control techniques and consider their test impacts.

Input Vector Control (IVC): For each logic gate, the quiescent current depends on its input combinations. Consider a three-input CMOS NAND gate as an example. For the "111" input combination, the three NMOS transistors are turned on and act as a short circuit; the gate's leakage current is the sum of the leakage current through the three PMOS transistors. For the "001," "010," "100," and "000" combinations, there are at least two NMOS transistors that are turned off in the pull-down network. In these cases, the "off" transistor on top of the stack has a positive source voltage, V_S.

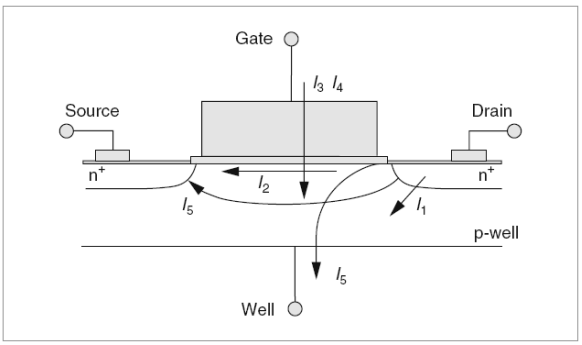


Figure 1: Leakage current mechanisms in deep-submicron transistors.

In the quiescent state, the leakage currents through all the transistors are equal. So, we can consider only the first "off" transistor on top in the pull-down tree as pertinent to our analysis. A positive V_S means a negative V_{GS}, which greatly reduces the leakage. A positive V_S also indicates the existence of body effect and a reduction in V_{DS}. Both effects increase the threshold voltage leading to further reduction in leakage.

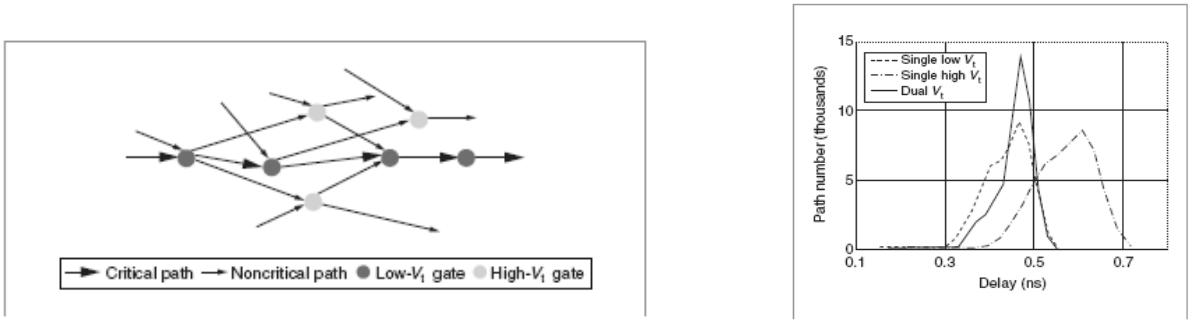
Since circuit's total leakage current depends on its primary inputs, applying the best input vectors to some circuits can cause the leakage current to decrease significantly [3]. Because of the exponential complexity with respect to the number of primary inputs, efficient algorithms to determine near-optimal solution based on random search or genetic algorithm have been developed [35]. Investigation shows that for a reasonably complex circuit, input vector control can result in about 30-35% saving in standby leakage using proper selection of input vector. This reduction in background leakage can improve the effectiveness of I_{DDQ} testing, particularly for testing complex circuits e.g. SoC, where all modules on chip except the one being tested can be applied the best vector for leakage reduction. Note that IVC may require hard-wiring the best input vector in the first level logic gates of a logic block or control point insertion [4]. Proper functioning of this extra logic needs to be checked during test while ensuring that it does not affect normal functionality.

Dual-V_t Design: For a logic circuit, we can assign a higher threshold voltage to some transistors in non-critical paths to reduce leakage current, while maintaining performance by using low-threshold transistors in critical paths. Therefore, no additional leakage-control transistors are necessary and we can achieve both high performance and low power dissipation simultaneously. Fig. 2(a) illustrates the idea of a dual-V_t circuit. Fig. 2(b) shows the path distribution of dual- and single-V_t CMOS for a 32-bit adder. Dual-V_t CMOS has the same critical delay as a single-low-V_t CMOS circuit, but we can assign the transistors in non-critical paths a high V_t to reduce leakage power. Hence, this dual-threshold technique can effectively reduce leakage power during both standby and active modes without incurring delay or area overhead. Because it can reduce background leakage, it can be beneficial for I_{DDQ} testing.

Let us investigate the benefits of combining the dual-threshold CMOS design technique and a vector-control technique for I_{DDQ} testing. For simplicity, we map the benchmark circuits to a library containing NAND gates, NOR gates, and inverters. The supply voltage is 1 V, and the low threshold voltage is 0.2 V. Using the algorithm designed by Wei et al. [22], we can transform the single-low-V_t circuit to a dual-V_t circuit with the optimal value for a high threshold voltage. We can then use a random search to choose the best vector from 1,000 randomly generated vectors. Thus, we capture the benefit of the vector-control technique on I_{DDQ} testing of a dual-threshold circuit. Results indicate that, for some shorts, combining the dual-threshold voltage design and leakage-control techniques can increase the fault current ratio by a factor of more than 10.

Supply Gating: A more promising technique is to stack transistors, supplying V_{SS} or V_{DD} through another control transistor [3]. The additional transistor in the stack effectively "gates" the V_{SS} or V_{DD} line during idle mode of the circuit to save leakage power. A variant of this gating technique, called *Multi- Threshold CMOS* (MTCMOS) uses high-V_t gating transistor along with low-V_t core [3]. This fits particularly well with regular structures such as data paths, where the gating transistor can be easily shared. The additional gating transistor in the charging/discharging path is a performance issue. A shared gating transistor requires careful sizing such that it is wide enough to sustain worst-case switching condition with acceptable performance loss. Since in the sleep mode, some output nodes are floated (using the small leakage current to hold their states), noise immunity becomes a robustness concern. The circuits in the sleep mode become susceptible to coupling noise or other power-transient events. Test engineers must face the challenge of deciding how to test the noise margin and as well as the worst-case delay overhead due to the gating transistor.

Shannon Cofactoring Based Dynamic Supply Gating: Low-leakage circuit design technique can directly help in improving I_{DDQ} testability. However, leakage control techniques based on transistor stacking that target active leakage reduction in logic circuits can also improve test power and test time. In [27], a circuit synthesis approach is proposed that can result low active power dissipation, while enhancing test cost and test confidence. The synthesis technique is based on structural transformation of a design using *Shannon's decomposition* and supply gating. It is observed that tree structure of a logic circuit due to Shannon's decomposition makes it intrinsically more testable than conventionally synthesized circuit, while at the same time entailing an improvement in active power. Significant improvement can be observed in three aspects of testability of a circuit: a) I_{DDQ} test sensitivity, b) test power during scan-based testing, and c) test



(a) (b)

Figure 2: a) Dual-threshold-voltage CMOS circuit; b) path distribution of dual- and single-Vt CMOS.

length (for both ATPG-generated deterministic and random patterns) [27].

Leakage Control in Memory: Leakage from embedded memory cells constitute a major part of system static power, typically in high-performance computing systems such as processor, SoC etc. which requires large on-board memory. The *de-facto* standard of embedded memory design is 6-transistor static random access memory. Leakage saving techniques in memory are predominantly based on variants of supply gating technique. A common scheme is *source biasing* [3] that applies "gating" at the source terminal of the NMOS transistors and applies a fixed bias at the virtual ground node to ensure data retention.

Leakage reduction techniques in memory will have positive impact on static current testing as well as on burn-in. In [31], improvement in I_{DDQ} testability for a GND-gating scheme applied to SRAM cells is proposed. During test mode, idle (not accessed) parts of the memory are "gated" using the most significant bits of the address line as gating control. Supply gating and source biasing techniques for memory, however, introduce new test challenges. A source-biased memory will have two distinct states (normal and supply-gated) and desired behavior in each state need to be checked during test. While read/write and access time failures need to be validated with the gating transistor "on" (normal mode of operation), the primary concern in the power saving mode (gating transistor "off") is data retention in memory cells. Test engineers need to ensure that the bias voltage is large enough to retain stored content in power-gated cells.

Thermal Stability during Burn-In: Leakage is a major issue during burn-in test, which is used to detect infant mortality types of defects. Leakage power is a dominating component of total power dissipation during burn-in test condition due to applied high supply voltage and temperature. In scaled technologies, during burn-in there is an exponential increase in junction temperature due to drastic increase in leakage power, higher transistor density and increase in die-to-package thermal resistance. An effective solution to the problem is to design a negative feedback system to stabilize the junction temperature by controlling the leakage power of a chip dynamically. In [29], such a system is proposed that continuously monitors the junction temperature and compares it with the target burn-in temperature. If the junction temperature is higher (lower) than the target temperature, the system decreases (increases) leakage current by decreasing (increasing) the reverse body bias of the chip.

2.3 Dynamic Power and Thermal Management

With technology scaling, active power per switching reduces due to scaling of V_{DD} and switching capacitance. However, faster clock and increasing device integration causes significant rise in overall dynamic power. The increase in dynamic power manifests as increased power density (computed as power dissipated per unit area) of the chip. Higher power density translates to higher junction temperature in the device layer, giving rise to localized "hotspots" due to limited cooling capacity of the packages. The power density for high performance microprocessors has been reported to be of the order of 50W/cm^2 for 100nm technology and is increasing further with scaling. Interestingly, localized hotspots are also a leakage concern, since the static power, in particular, the subthreshold leakage component increases exponentially with temperature [29], potentially causing thermal runaway condition. It has been almost mandatory to incorporate power reduction techniques in nanoscale CMOS designs to reduce average power dissipation and avoid temperature-induced reliability concerns as well. Next, we will discuss some major dynamic power reduction techniques and associated test impact.

Circuit optimization for low power: Circuit-level design techniques for dynamic power reduction typically include downsizing logic gates (in order to reduce effective switching capacitance) [5] and static assignment of multiple threshold [22] or multiple supply voltages [12]. These techniques essentially exploits the timing slack available in the shorter paths and make them slower, effectively equalizing the timing paths. The undesirable effect of this optimization on test is large increase in critical delay paths, which complicates the path selection process for delay testing and speed binning. This also becomes a major reason for yield loss due to parameter variations.

Clock gating: Clock gating is an effective low-overhead technique for reducing power in the clock line by shutting off clock switching in the idle logic blocks. Conditionally switching the clock line in a localized manner, however, causes test concern. Clock gating increases temporal variations in supply current drawn from the power-grid (which can be modeled as a big RLC network) causing inductive voltage droop. Such local transient fluctuations in power-grid affect signal propagation through logic gates resulting in timing failure unless sufficient margin is not maintained during design time. Delay test generation and application require mimicking the worst-case droop in power grid to realistically capture the delay variation.

Advanced Power and Thermal Management: Due to quadratic dependence of dynamic power of a circuit on its operating voltage, supply voltage scaling (along with commensurate scaling of operating frequency) has been extremely effective in reducing the power dissipation. As we have noted earlier, high performance systems such as processors or SoCs, also suffer from high power density issue that results in high junction temperature. The temperature issue is typically addressed by monitoring the temperature of processing units (using distributed temperature sensors) and throttling clock frequency or reducing voltage when the temperature goes beyond a threshold [34].

Such dynamic power and thermal management techniques are attractive since they can achieve maximum performance under a power-temperature envelope. However, they can have undesirable consequences on test. Circuit delay changes in a non-linear fashion with voltage and temperature and the dependence of delay on operating condition changes

unpredictably with process variations. Moreover, temperature-induced variations are often local due to presence of localized thermal gradient. This makes a static design-time delay calibration at different operating conditions very unrealistic. However, an important test challenge is to define the worst-case timing condition during test. Different activity levels in different parts of a dies cause variations in junction temperature. The worst-case condition may correspond to a non-uniform power level, which may be difficult to emulate in test mode using an ATE. Testing all the processing units for the worst-case condition may cause over-testing leading to yield loss. On the other hand, leaving some paths untested under worst-case temperature distribution may result test escape. Finally, during functional testing, an ATE need to correctly predict the thermal trigger point in order to avoid false alarm. The problem aggravates for emerging multi-core platforms that distribute workload (with the help of operating system) among multiple cores to achieve power efficiency. Since the thermal conditions on different cores are functions of applications and operating system, it is difficult to structure delay test for the worst-case thermal distribution.

3. Test Considerations under Process Variations

Process variations can cause parametric yield loss since a circuit designed at nominal process corner may fail to satisfy target delay, leakage or noise margin under parameter variations [11-14, 18]. Conventional wisdom dictates a conservative design approach (e.g., scaling up the V_{DD} or upsizing logic gates) to avoid large number of chip failures due to variations. However, such techniques come at the cost of considerable increase in power and/or die area. Over the past few years, researchers have looked for alternative design paradigms to ensure parametric yield under variations with minimal design overhead. Design approaches that are robust with respect to process variations and at the same time, suitable for low power operation have also been investigated. Static and dynamic voltage scaling as well as multi-threshold approach to reduce power dissipation, however, requires comfortable timing margin built into the design. Process variations make low-power design difficult since the timing margin becomes a probabilistic parameter. Further, design optimization using voltage scaling, dual-V_{th} assignment and gate sizing to improve power typically increase the number of critical paths in a circuit, giving rise to the so-called "wall effect". This causes concern in terms of path delay fault testability and parametric yield. In order to appropriately capture the impact of parameter variations in circuit functionality and address the associated design and test challenges, a paradigm shift from static to statistical analysis/design approach seems more appropriate.

3.1 Statistical Analysis and Design

With inter-die and intra-die parameter variations, traditional approach to static analysis and design of circuit considering specific target frequencies and power budgets (dynamic + leakage power) in mind becomes less effective. Considering a V_{th} distribution as shown in Fig. 3(a), circuit delay and power also can be modeled as a statistical

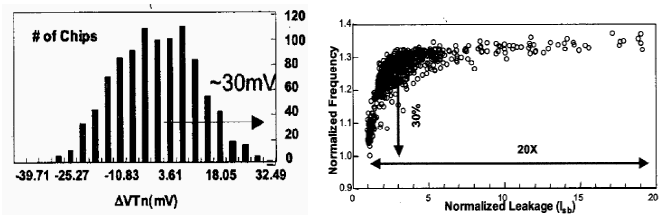


Figure 3: a) Example of die-to-die V_{th} variations for 180nm CMOS process [2]; b) leakage and frequency variations of a high-performance design [2].

distribution using correlated model of gate and interconnect delays. In recent years, statistical analysis of timing (SSTA) and power has been extensively explored [6, 8, 11]. Interestingly, when a circuit parameter such as delay follows a statistical distribution, we can meet a predefined delay target only with certain probability. This probability, referred as parametric yield, can be estimated by modeling the distribution as a probability density function (typically Gaussian) and computing the probability to meet a pre-defined target. Several parametric yield models have been proposed to consider impact of various sources of variations on circuit delay and power [4, 18]. On the other hand, gate sizing and/or V_{th} assignment have been primarily used as a tool to modulate the circuit delay distribution for yield improvement or yield-constrained area/power minimization [5, 9, 12].

Statistical delay-defect simulation and test: Variability of speed paths due to process variations raises questions regarding the effectiveness of conventional structural delay testing. Transition delay testing that model delay defects as large gate delay fault, provides a simple and affordable delay testing methodology. However, from the test viewpoint, there are several relevant questions. How effective is transition delay fault model in nanoscale CMOS designs? Since small delay defects on short paths are undetectable in transition delay test, high transition fault coverage often does not guarantee high delay fault coverage for nanometer designs [30]. Augmenting the transition fault test with path delay testing for a number of selected critical paths is becoming almost mandatory. However, path delay testing under parameter variations raises some important questions. What will be a realistic target for path delay coverage? How to measure this coverage reliably? How to select the critical paths? Can conventional static timing analysis predict them?

Spatial and temporal parameter variations are clearly major issues in path selection and determination of delay fault coverage. Conventional static timing analysis that model delay as a single static value using worst-case condition is not effective for path selection under variability, since the actual critical paths in different chip instances may differ. In order to estimate good delay test coverage and avoid yield loss, we need to accurately determine the sources of device parameter variations (such as lithography variables, chemical mechanical polish etc.); estimate their impact on circuit delay and model circuit and path delay as statistical distributions. Once the path delay distributions are determined, a set of critical paths can be selected for delay testing based on the probability of the paths to meet a delay target [25]. The

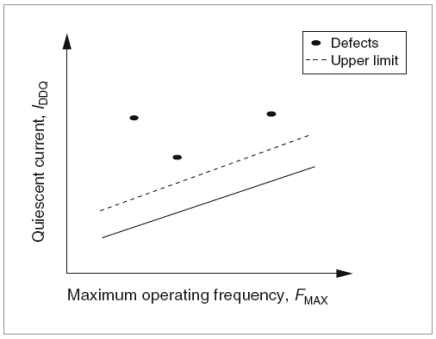


Figure 4: Quiescent current (I_{DDQ}) versus maximum operating frequency (F_{MAX}).

optimal selection of paths will depend on the delay defect model and the required test confidence level. Moreover, diagnosis of delay defects often requires different set of paths from those required for path delay test [26]. The task of path selection needs to be combined with pattern generation to sensitize the paths for worst-case timing condition. Note that under statistical delay defect model, path sensitization becomes probabilistic. Therefore, developing ATPG tools that employ statistical timing constraints during the justification process is a challenging problem.

Leakage Variations: The variation in transistor parameters in scaled technologies not only causes delay variations but also significant spread in I_{DDQ} distribution [18]. Fig. 3(b) shows that leakage variations for a present generation processors can be 20X for a frequency variation of 30%. High-performance chips typically require satisfying a leakage bound. Hence, a leakage binning process similar to speed binning is employed to weed out chips failing leakage specification. Clearly, increased variations in leakage result in yield loss. In dynamic circuits, such as register file, leakage variation increases the number of dies with insufficient robustness. Further, to avoid test escape, we need to apply the right patterns during test to excite the worst-case leakage condition.

I_{DDQ} Test under Leakage and Delay Variations: It is interesting to note that we can exploit the intrinsic dependencies of transistor and circuit leakage on clock frequency, temperature, and optimal body bias to distinguish between fast and defective ICs [23]. Transistor and circuit parameters can be correlated to achieve leakage-based testing solutions with improved sensitivity. Because of short-channel effect (SCE), reducing transistor effective channel length (L_{eff}) lowers the threshold voltage. Consequently, it increases I_{OFF} and I_{DDQ} but provides a higher maximum operating frequency (F_{MAX}). At high F_{MAX}, the high leakage might stem from the shorter channel length and lower threshold voltage. But at low F_{MAX}, the high leakage is most likely due to the faulty circuit. Therefore, a fixed I_{DDQ} test limit is not enough for low-voltage-scaled circuits. We can set an adjustable test limit based on I_{DDQ} and F_{MAX} to establish a two-parameter test limit that distinguishes fast and slow die from those that are defective. Fig. 4 illustrates the relationship between quiescent current and maximum frequency. The I_{DDQ} limit climbs as F_{MAX} increases. The defect dots show higher leakage than the proposed adjustable I_{DDQ} limit shown by the dashed line. When I_{DDQ} is high, F_{MAX} should be high, representing a fast IC. However, if F_{MAX} happens to be low when I_{DDQ} is high, the IC is most likely defective, and this situation is a prime candidate for failure analysis to determine what defects are actually present. The two-parameter test technique improves the effectiveness of I_{DDQ} testing for single-threshold, fast CMOS ICs. This test method was improved further to improve its sensitivity, using temperature or body bias.

Improvement in Leakage Yield: Interestingly, low leakage design techniques sometimes help to reduce the leakage spread too, thus improving the leakage yield. For example, the Shannon decomposition based supply gating technique [27] (SBS) discussed in Section 2 can also improve the parametric yield due to leakage variability. Since the SBS can reduce the maximum I_{DDQ} significantly even in fast process corner, it is possible to salvage previously failing chips. Similarly, reverse body biasing technique for improving leakage can also help in increasing leakage yield by squeezing the leakage distribution [15].

3.2 Avoidance of Variation-Induced Failures

A new paradigm to design low-power circuits under process variation, referred as *CRISTA* [19], is based on the concept of avoiding delay failures with clever synthesis. *CRISTA* makes a circuit amenable to aggressive voltage scaling, while being robust to parametric failures by using a synthesis technique that 1) isolates and predicts the set of possible paths that may become critical under process variations, 2) ensures they are activated rarely, and 3) tolerates any delay failures in the set of critical paths by adaptively operating in two-cycles (assuming all standard operations are single cycle). It employs the notion of *critical path isolation* that indicates confinement of critical paths of a design to few known logic blocks. This is accomplished by partitioning a circuit into multiple cofactors using Shannon decomposition and then using gate-sizing to create appropriate timing margins between cofactors. Any

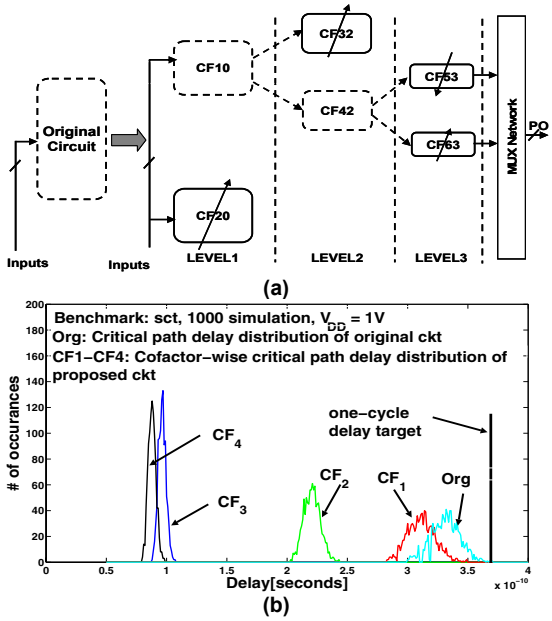


Figure 5: a) Hierarchical expansion and sizing of cofactors resulting in b) desired path delay distribution for failure avoidance (shown for a MCNC benchmark) [19].

delay errors (that may occur under a single cycle operation) are predicted dynamically by decoding a small set of inputs and are avoided with two-cycle operations. Fig. 5(a) shows the partitioning and gate sizing steps in the synthesis process and Fig. 5(b) plots the skewed path delay distribution of critical and non-critical blocks of an MCNC benchmark after the synthesis step. Simulation results show that the technique can achieve about 60% average power saving under yield constraint at moderate area overhead [19].

3.3 SRAM Parametric Failures under Variations

Die-to-die and within-die variations in process parameters result in the mismatch in the strengths of similar transistors in an SRAM cell, which causes functional failures degrading memory yield [14]. Conventionally, redundant rows/columns are used in memories to improve yield. These redundancy techniques have limitation on the number of faulty rows/columns it can handle, due to resource limitation and design complexity. In particular, the failures due to within-die variations are randomly distributed across the dies, resulting in a large number of faulty rows/columns. Recovery from such defects is difficult to handle by row/column redundancy alone. Moreover, SRAM failures due to process variation change depending on operating condition (e.g. supply voltage, frequency, temperature). The operating condition changes dynamically, which makes static testing and repairing techniques less effective. Error Correcting Codes (ECC), employed to correct transient faults (such as soft error) in memory, can also be used to correct failures due to process variations. However, ECC has limitations on the number of error bits it can correct and incurs considerable hardware overhead.

Fig. 6 summarizes the failure mechanisms in SRAM under process variations and shows mapping of the failures to the logic fault models [28]. Among the failure mechanisms: 1) *SRAM access failures and SA functional failures* may show themselves as either incorrect read faults or random read faults, depending on the noise level and the sense-amplifier offset voltage; 2) *SRAM flipping read failure* is modeled by read destructive fault or deceptive read destructive fault, based on the time the cell flips and how fast bit-lines responds to that flip; and 3) *SRAM hold failure* is modeled as data retention fault if the failure occurs at the nominal supply voltage. However, most of the hold failures happen in the standby mode (when the supply voltage is decreased to reduce leakage power) [14]. Extending the concept of data retention fault, a new fault model named *low supply data retention fault* is developed to describe flipping failures occurring due to application of low supply voltage in the standby mode [28].

By mapping the process variation related failures to logic fault models, memory test can be designed to target failures in nanoscale SRAMs. The flipping read failure in SRAM cells has a high probability of occurrence (about 2% of cells). Most of the flipping read failures show themselves as deceptive read destructive faults, which is overlooked in conventional memory test. Low supply data retention faults are also likely to occur (about 4%). These fault types are not emphasized by the conventional March tests. Moreover, it is difficult to optimize conventional March test to cover the process-induced failures without trading off test time [28]. Novel DFT circuits can help in reducing the test time without affecting fault coverage. One such technique, referred as *double sensing*, is applied to test the read stability of SRAM cells. The idea is to replace the consecutive read operations for detecting deceptive read destructive faults by parallel sense amplifiers to sample the bit-lines twice during one read cycle [28]. The technique is effective to reduce test time and additionally, can be used during normal operation for online detection and correction of random read failures.

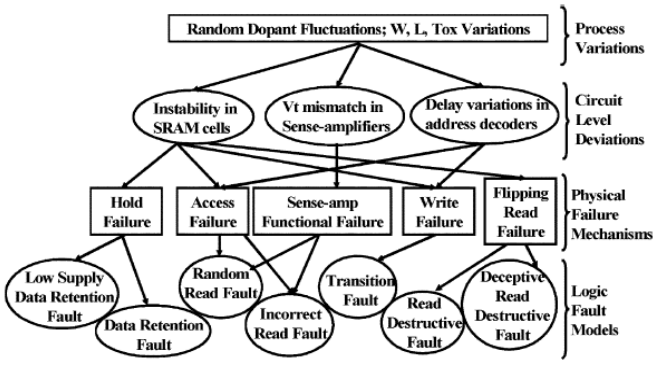


Figure 6: Process variation induced failure mechanisms and logic fault models in SRAM [28].

4. Self-Calibration and Self-Repair

Post-silicon strategies for self-calibration and self-repair constitute a promising class of solutions to address power and variation induced test challenges. Below, we discuss some important calibration and repair schemes for logic and memory circuits that can simplify the test procedure and reduce test cost with moderate design overhead.

4.1 Self-Calibration and Repair in Logic Circuit

As discussed earlier, process variation in logic circuit primarily manifest as variations in delay, leakage and noise margin. The shift in circuit parameters can be detected using on-chip process sensor and deviation in parameters due to variations can be compensated by appropriate technique.

RAZOR: One such technique, called *RAZOR,* uses dynamic detection and correction of circuit timing errors to adjust the supply voltage [21]. It potentially eliminates the requirement of delay margin during design phase. Razor relies on a combination of architectural and circuit-level techniques for

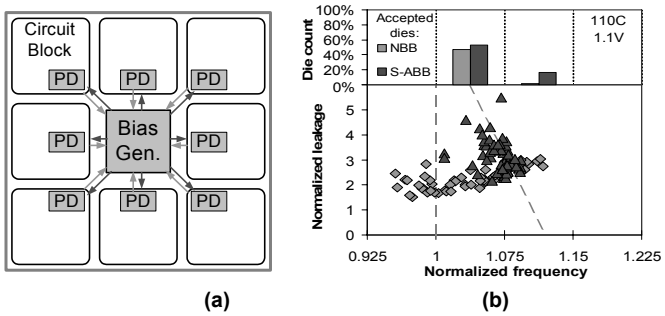


Figure 7: a) Adaptive body biasing scheme considering within-die delay variations; b) leakage vs. frequency distribution of an adaptive body biasing scheme that considers both inter- and intra-die variations [15].

efficient detection and correction of delay path failures by using a shadow latch controlled by a delayed clock corresponding to each critical flip-flop. In a given clock cycle, if the combinational logic meets the timing requirement for the main flip-flop, then it writes the correct data. On the other hand, if the combinational logic does not complete computation in time, the main flip-flop will latch an incorrect value, while the shadow latch will write the late-arriving correct value. A simple correction scheme restores the correct value from the shadow latch. Such an adaptive technique definitely helps to address the uncertainty in path delay due to variations reducing cost for delay test and speed binning.

Body biasing and Effect on Delay Test: Body bias has strong impact on leakage and performance of a die and thus has been investigated as a potent process adjustment tool. While forward body bias (FBB) helps to improve performance in active mode (by lowering the V_{th}), reverse body bias (RBB) is effective to reduce leakage power (by increasing the V_{th}). A practical application of body bias to adjust process variations requires accurate detection of process shift at different parts of a circuit and application of an optimal body bias voltage, which maximizes the performance under leakage constraint. Typically, on-chip process sensors for delay or leakage monitoring are used to determine the process shift during test. In [15], a bidirectional adaptive body bias (ABB) technique, shown in Fig. 7(a) is used to compensate for die-to-die parameter variations by applying optimum pMOS and nMOS body bias voltages to each die. To account for intra-die variations, an enhancement of this technique is proposed that requires a phase detector (PD) to determine frequency of a block from its critical path replica. The central bias generator considers output of all PDs to determine the optimal bias. Measurement results show that the technique results in an increase in number of acceptable dies as well as number of high-frequency dies (Fig. 7(b)).

An ABB technique effectively reduces the delay spread in each chip, thereby improving path delay testability. An investigation was performed in [33] to observe the impact of body biasing on delay fault test under both inter and intra-die process variations. Simulation results show that with a fixed optimum forward body bias one can considerably reduce the delay fault test overhead due to process parameter variations. However, with the adaptive body biasing technique one requires to test only a few paths for delay faults, while achieving very high test quality.

Process Compensation in Dynamic Circuits: Increasing I_{OFF} with process scaling has forced designers to upsize the keeper in dynamic circuits to obtain an acceptable robustness under worst-case leakage conditions. However, large (over 20x) variation in die-to-die NMOS I_{OFF} indicates that 1) a large number of low leakage dies suffer from the performance loss due to an unnecessarily strong keeper, while 2) the excess leakage dies still cannot meet the robustness requirements with a keeper sized for the fast corner leakage. A process-Compensating Dynamic (PCD) circuit technique that improves robustness and delay variation spread by restoring robustness of worst-case leakage dies and improving

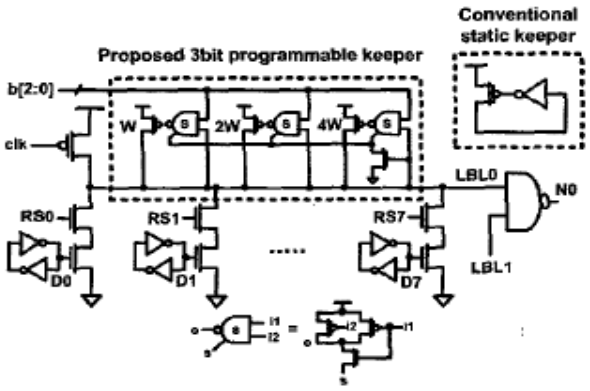


Figure 8: Register file with process compensating dynamic circuit technique (digitally programmable keeper size can be configured to be: 0, W, 2W, … , 7W) [16].

performance of low-leakage dies is presented in [16]. Fig. 8 shows the PCD scheme with a digitally programmable 3-bit keeper applied on an 8-way register file local bitline (LBL). Such a keeper enables 10% faster performance, 35% reduction in delay variation, and 5x reduction in robustness failing dies over conventional static keeper design in 90nm dual-V_{th} CMOS process [16]. As before, effectiveness of the compensation scheme largely depends on efficient process detection mechanism. Together, they can be very effective to improve test cost and yield for dynamic circuits.

Delay Calibration: The wide variation in operating frequency (e.g. ~30% in a processor) has introduced the concept of frequency or speed binning. Speed-binning requires calibration of maximum operating frequency (F_{max}) at different operating conditions such as supply voltage, temperature etc. In the simplest scenario, it is desired to determine F_{max} corresponding to a given operating voltage under worst-case temperature condition. The process is expensive in terms of both test application time and complexity of the test hardware since it requires testing at multiple frequencies for a given supply voltage. Consequently, test cost associated with speed binning is significant. The situation becomes worse when it is required to calibrate F_{max} at multiple operating voltages. Calibration of F_{max} at different operating voltages is required for primarily two reasons: (a) in a Dynamic Voltage and Frequency Scaling (DVFS) system [21], the adaptation hardware requires to apply correct operating frequency corresponding to a scaled supply; and (b) to sort chips in correct voltage-frequency (V-F_{max}) bins, so that chips at different bins can be used for different applications. It has been observed that frequency vs. voltage relationship not only changes from chip to chip but changes in an unpredictable manner at different voltage points for the same chip as well. Thus a static design-time calibration cannot provide practical solution [32].

Given the complexity and cost of speed binning at just one voltage, it is important to develop design techniques to aid the binning process based on structural testing. Earlier it has been demonstrated that speed binning using structural delay testing correlates well with binning process based on functional tests. Conventional approach based on creating a critical-path replica cannot reliably represent the delay of the actual critical path due to local within-die variations. In order

to measure the frequency shift accurately, it is better to consider the actual timing paths in the circuit. In [32], a low-overhead design solution for characterizing F_{max} of a circuit at different operating voltages is presented. The basic idea is to choose a small set of representative paths in a circuit based on their voltage sensitivity and dynamically configure them into ring oscillator to compute F_{max}. The proposed calibration mechanism is all-digital, robust with respect to parameter variations, reasonably accurate (with an average error of 2.8% for ISCAS89 benchmarks) and incorporates minimal hardware overhead.

4.2 Self-Repairing SRAM

With the limitations of the existing fault-tolerant techniques, SRAM that can repair itself and reduce the number of failures would be very effective for memory yield improvement. Next, we will discuss two major self-repair techniques for SRAM at the circuit and architecture level.

Adaptive Body Biasing: A V_{th} shift toward low V_{th} process corners, due to inter-die variation, increases the read and the hold failures of SRAMs. This is because of the fact that, lowering the V_{th} of the cell transistors increases V_{READ} and V_{TRIPRD}, thereby increasing read failures [14]. The negative V_{th} shift increases the leakage through the transistor N_L, thereby, increasing the hold failures. On the other hand, for SRAM arrays in the high-V_{th} process corners, the probabilities of access failures and write failures are high. This is principally due to the reduction in the current drive of the access transistors. The hold failure also increases at the high V_{th} corners, as the trip-point of the inverter PR-NR increases with positive V_{th} shift. Hence, the overall cell failure increases both at low and high-V_{th} corners and is minimum for arrays in the nominal corner. Consequently, the probability of memory failure is high at both low-V_{th} and high-V_{th} process corners.

Let us now discuss the effect of the body-bias (applied only to NMOS) on different types of failures. Application of reverse body-bias increases the V_{th} of the transistors which reduces V_{READ} and increases V_{TRIPRD}, resulting in a reduction in the read failure [14, 17]. The V_{th} increase due to RBB also reduces the leakage through the NMOS thereby reducing hold failures. However, an increase in the V_{th} of the access transistors due to RBB increases the access and the write failures. On the other hand, application of FBB reduces the V_{th} of the access transistor, which reduces both access and write failures. However, it increases the read (V_{READ} increase and V_{TRIPRD} reduces) and hold (leakage through NMOS increases) failures [17].

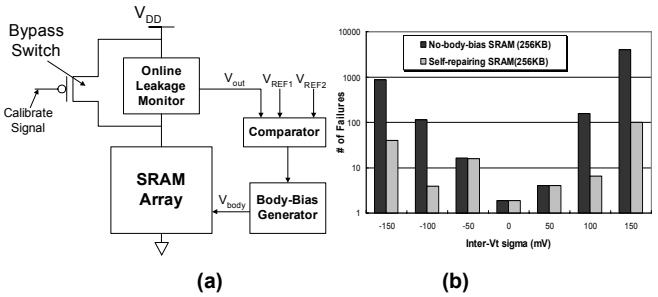


Figure 9: a) Self-repairing SRAM scheme; b) reduction in number of failures in 256KB memory array [20].

To determine the correct body bias to apply to the SRAM chip for failure probability improvement, the process corner, in which the memory chip sits, needs to be determined. An effective way to perform V_{th} binning is to use leakage monitoring. The random intra-die variation in threshold voltage results in significant variation in cell leakage, particularly, the sub-threshold leakage. In a self-repairing SRAM using "Leakage Monitoring", the measured leakage is compared with the reference currents to identify the inter-die process corner of the chip. Based on this measurement, the right body bias can be applied to the chip. The schematic of a self-repairing SRAM array with self-adjustable body-bias generator is shown in Fig. 9(a) [20]. Experimental results on reduction in number of failures shown in Fig. 9(b) appear promising to contain process-induced failures in SRAM.

Adaptive Remapping in Cache: An architecture-level technique proposed in [10] detects and replaces faulty cells by adaptively remapping the cache. This architecture assumes that the cache is equipped with a built-in-self-test (BIST) unit, which tests the entire cache and detects faulty cells due to parameter variations. Fig. 10 shows the anatomy of the process-tolerant cache architecture [10]. The scheme downsizes the cache by forcing the column MUX to select a non-faulty block in the same row if the accessed block is faulty. It maps the whole memory address space into a resized cache such that the remapping is transparent to the processor.

Conventionally, a cache is divided into cache blocks (e.g. 256 bit per block, Fig. 10) and several cache blocks are stored in a single row. A block is considered faulty even if a single cell in a block is faulty. BIST detects the faulty blocks and feeds the information into the configuration storage.

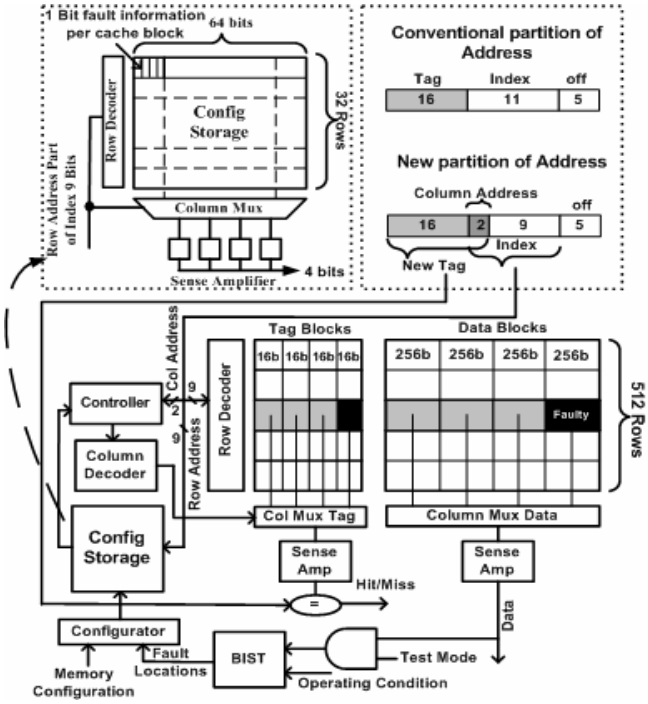


Figure 10: Architecture of 64K process-tolerant cache [10].

Configuration storage is a small memory, which stores 1-bit fault information per cache block. In conventional cache access, all the blocks in a cache row are selected simultaneously by a single wordline. Finally, column MUX chooses one block based on column address. The key idea of the proposed architecture is to force the column MUX to select another block in the same row, if the accessed block is faulty. Based on the fault information stored in the *config* block, a controller forces the column MUX to select another non-faulty block (if the accessed block is faulty) in the same row by altering the column address, effectively resizing the cache. In case of a faulty block access, this scheme selects the first available non-faulty block. Hence, as long as there is one non-faulty block in each row, this architecture can correct any number of faults. The cache access time remains unchanged, although there is a performance hit due to increased miss rate at diminished cache size. However, the yield improves by up to 94% compared to a yield improvement of 34% using redundant rows/columns at the cost of minimal area and power overhead.

5. Conclusions

In this paper, we consider the growing impact of power and process variations in nanoscale design and their impact on manufacturing test and yield. New failure mechanisms in logic circuits and SRAM have emerged due to inter- and intra-die process parameter variations. Hence, new test methodologies are required. However, the test problem becomes more complicated with new design methodologies/paradigms being adopted to cope with power and variation problems. Existing techniques on testing logic and memory circuits, fault diagnosis and fault tolerance may not work well under the new low-power statistical design environment. Besides, circuit and architectural techniques for low-power and variation-tolerant design often imposes conflicting requirements on test resulting in increased test complexity and cost. We believe that designers need to consider testability and yield in the design optimization framework in order to limit the growing test complexity and test cost as well as to achieve higher test confidence. Self-calibration and self-repair techniques appear promising to reduce test-cost, however, design overhead associated with these techniques should be minimized.

6. References

[1] International Technology Roadmap for Semiconductors, *http://public.itrs.net*, 2004.

[2] S. Borkar et al., "Parameter Variations and Impact on Circuits and Micro-architecture", *DAC,* 2003, pp. 338-342.

[3] K. Roy, S. Mukhopadhaya, and H. Mahmoodi-Meimand, "Leakage Current Mechanisms and Leakage Reduction Techniques in Deep-Submicrometer CMOS Circuits," *Proc. IEEE*, 2003, pp. 305-327.

[4] Lin Yuan and Gang Qu, "A combined gate replacement and input vector control approach for leakage current reduction", *IEEE TVLSI*, 2006.

[5] E. T. A. F. Jacobs and M. R. C. M. Berkelaar, "Gate Sizing Using a Statistical Delay Model", *DATE,* 2000, pp. 283-290.

[6] K. Kang et al., "Statistical Timing Analysis using Levelized Covariance Propagation", *DATE*, 2005, pp. 764-769.

[7] A. Agarwal et al., "Circuit Optimization using Statistical Timing Analysis", *DAC,* 2005, pp. 321-324.

[8] H. Chang and S. S. Sapatnekar, "Statistical Timing Analysis Considering Spatial Correlations using a Single PERT-like Traversal", *ICCAD*, 2003.

[9] M. Mani et al., "An Efficient Algorithm for Statistical Minimization of Total Power under Timing Yield Constraints", *DAC*, 2005, pp. 309-314.

[10] A. Agarwal et al., "A Process-Tolerant Cache Architecture for Improved Yield in Nanoscale Technologies", *IEEE TVLSI*, 2005, pp. 27-38.

[11] M. C.-T. Chao et al., "Static Statistical Timing Analysis for Latch-Based Pipelined Designs", *ICCAD 2004*, pp. 468-472.

[12] A. Srivastava and D. Sylvester, "A General Framework for Probabilistic Low-Power Design Space Exploration considering Process variation", *ICCAD 2004,* pp. 808-813.

[13] A. Bhavnagarwala et al., "The impact of intrinsic device fluctuations on CMOS SRAM cell stability," *IEEE JSSC*, pp. 658-665, April 2001.

[14] S. Mukhopadhyay et al, "Statistical design and optimization of SRAM for yield enhancement," *ICCAD*, 2004, pp. 10-13, 2004.

[15] J.W. Tschanz et al, "Adaptive body bias for reducing impacts of die-to-die and within-die parameter variations on microprocessor frequency and leakage," *IEEE JSSC*, 2002, pp. 1396-1402.

[16] C. H. Kim et al., "A Process Variation Compensating Technique for Sub-90nm Dynamic Circuits", *Symp. on VLSI Circuits*, 2003.

[17] S. Mukhopadhyay et al, "Modeling and estimation of failure probability due to parameter variations in nano-scale SRAMs for yield enhancement," *Symp. on VLSI Circuits*, 2004, pp. 64 – 67.

[18] R. Rao, et al, "Parametric yield estimation considering leakage variability," *DAC*, 2004, pp. 442 - 447.

[19] S. Ghosh et al, "A New Paradigm for Low-power, Variation-Tolerant and Adaptive Circuit Synthesis Using Critical Path Isolation," *ICCAD*, 2006.

[20] S. Mukhopadhyay et al, "Reliable and self-repairing sram in nano-scale technologies using leakage and delay monitoring," *ITC*, 2005.

[21] D. Arnst et al., "Razor: A low-power pipeline based on circuit-level timing speculation," *Proc. MICRO-36*, 2003, pp. 7–18.

[22] L. Wei et al., "Design and Optimization of Dual Threshold Circuits for Low-Voltage Low-Power Applications," *IEEE TVLSI*, 1999.

[23] A. Keshavarzi et al., "Multiple-Parameter CMOS IC Testing with Increased Sensitivity for IDDQ," *ITC*, 2000, pp. 1051-1059.

[24] K.-T. Cheng et al., "Test Challenges for Deep Sub-Micron Technologies," *DAC*, 2000, pp. 142-149.

[25] J.-J. Liou et al., "False-Path-Aware Statistical Timing Analysis and Efficient Path Selection for Delay Testing and Timing Validation," Proc. *DAC*, 2002, pp. 566-569.

[26] A. Krstic et al., "Enhancing Diagnosis Resolution for Delay Defects Based upon Statistical Timing and Statistical Fault Models," *DAC*, 2003.

[27] S. Ghosh et al., "Shannon Expansion Based Supply-Gated Logic for Improved Power and Testability", *ATS*, 2005.

[28] Q. Chen et al., "Efficient Testing of SRAM With Optimized March Sequences and a Novel DFT Technique for Emerging Failures Due to Process Variations", *IEEE TVLSI*, 2005.

[29] M. Meterelliyoz et al., "A Leakage Control System for Thermal Stability During Burn-In Test", *ITC* , 2005.

[30] T.M. Mak et al. "New Challenges in Delay Testing of Nanometer, Multigigahertz Designs", *IEEE Design & Test of Computers*, 2004.

[31] S. Bhunia et al., "A High Performance I_{DDQ} Testable Cache for Scaled CMOS Technologies", *ATS*, 2002.

[32] S. Paul et al., "Low-Overhead Design Technique for Calibration of Maximum Frequency at Multiple Operating Points", to appear, *ICCAD* 2007.

[33] B. C. Paul et al., "Impact of Body Bias on Delay Fault Testing of Nanoscale CMOS Circuits," *ITC*, 2004.

[34] R. McGowen et al., "Power and Temperature Control on a 90-nm Itanium Family Processor", *IEEE JSSC*, 2006.

[35] M.C. Johnson, D. Somasekhar, and K. Roy, "Models and algorithms for bounds on leakage in CMOS circuits," *IEEE Trans. Compur-Aided Des. Integr. Circuits Syst.*, vol. 18, no. 6, pp. 714-725, 1999.

[36] http://www.research.ibm.com/aceed/2005/proceedings/panel-anderson.pdf