Technology Variability from a Design Perspective

Borivoje Nikolić, Bastien Giraud, Zheng Guo[1], Liang-Teck Pang[2], Ji-Hoon Park, Seng Oon Toh
Electrical Engineering and Computer Sciences, University of California, Berkeley
[1]now with Intel Corp, Hillsboro, OR, [2]now with IBM T.J. Watson Research Center, Yorktown Heights, NY

Abstract—**Increased variability in semiconductor process technology and devices requires added margins in the design to guarantee the desired yield. Variability is characterized with respect to the distribution of its components, its spatial and temporal characteristics and its impact on specific circuit topologies. Approaches to variability characterization and modeling for digital logic and SRAM are reviewed in this paper. Transistor and ring oscillator arrays are designed to isolate specific systematic and random variability components in the design. Distributions of SRAM design margins are measured by padded-out cells and minimum operating voltages for the entire array. Correlations between various components of variability are essential for adding appropriate margins to the design.**

I. Introduction

Increasing process variability is perceived as one of the major roadblocks for continued technology scaling [1]. In sub-100nm technologies, it is becoming increasingly difficult for the device tolerances to track the scaling rate of the minimum feature sizes. Variations are generally characterized as within-die (WID), die-to-die (D2D) and wafer-to-wafer (W2W) [2]. While the W2W variations dominated in the past, with scaling of the technology, WID and D2D variations can occupy a majority of the process spread. Variation of process and device parameters can be systematic or random, spatially or temporally correlated. Sources of variability are in the transistors, interconnect, and in the operating environment (supply and temperature) [3]. Device parameters vary systematically because of deviations in nominal widths, lengths, film thicknesses and dose of implants due to the manufacturing process [4]. Random device parameter fluctuations are associated with atomistic variations in device structure.

Simultaneously, with process scaling, the nature of IC design has gradually shifted to become power limited. In current and future technology nodes, the optimization for energy consumption is as important as optimization for performance [5]. To sustain the current trend in technology scaling, which dictates higher parallelism in each technology generation, optimization for energy requires further lowering of the supply voltages. However, to mitigate the impact of increased variability, appropriate design margins have to be added to every component of an integrated circuit. In addition, the sensitivity of power and performance to process variations increases at low supply voltages. Therefore, the requirements for robust operation often contradict the needs for energy efficiency, and this is exacerbated by variability.

This paper reviews various classes of technology variability, analyzes their interactions, and presents methods for their accounting in the design margins. Characterization of variability is essential for setting the appropriate design margins. Numerous characterization structures have been developed that allow for collection of large datasets of process and device parameters, as well as their spatial and temporal characteristics and correlations. Structures for characterizing digital gates and SRAM have been designed to generate large datasets suitable for evaluating the distributions of device parameters and their impact on circuit yield.

II. Technology Variability

There are many sources of variability in the design and numerous ways to classify them. The primary sources of variability are the transistors, the interconnect, supply and temperature.

CMOS process parameter variability is often classified into three categories: known systematic, known random and unknown [6]. Systematic process variations are deterministic shifts in space and time of process parameters, whereas random variations change the performance of any individual instance in the design in an arbitrary way. Systematic variations are, in general, spatially correlated. In practice, although many of the systematic variations have a deterministic source, they are either not known at the design time, or are too complex to model, and are thus treated as random. As a result, many of the sources of variability are not modeled in the design kits and have to be treated as random in the design process. The resulting 'random' variation component, depending on the way systematic variability is modeled, will often appear to have a varying degree of spatial correlation [7].

Spatial tolerances in the manufacturing process are classified as WID, D2D, W2W and lot-to-lot (L2L) [2]. Variations reflect both the spatial as well as the temporal characteristics of the process and cause different dies and wafers to have different properties. The performance of the manufacturing equipment, expressed through the dose, speed, vibration, focus, or temperature, varies within one die and from die to die. Those parameters that vary rapidly over distances smaller than the dimension of a die result in WID variations whereas variations that change gradually over the wafer will cause D2D variations. Similarly, even more parameters vary from wafer to wafer (W2W variations) and between different manufacturing runs (L2L variations).

Many sources of systematic spatial variability can be attributed to the different steps of the manufacturing process. The photolithography and etching contribute significantly to variations in nominal lengths and widths due to the complexity required to fabricate sublithographic lines that are much narrower than the wavelength of light used to print them [8]. Significant contributors in this area include temperature non-

978-1-4244-5760-1/10/$26.00 ©2010 IEEE

uniformities in the critical post-exposure bake (PEB) and etch steps. Variation in film thicknesses (e.g., oxide thickness, gate stacks, wire and dielectric layer height) is due to the deposition and growth process, as well as the chemical-mechanical planarization (CMP) step. Additional electrical properties of CMOS devices are affected by variations in the dosage of implants, as well as the temperature of annealing steps. In recent technologies, overlay error, mask error, shift in wafer scan speed, rapid thermal anneal and the dependence of stress and proximity on layout have become notable sources of systematic variations.

Random device parameter fluctuations stem mainly from line-edge roughness (LER) [9], Si/SiO_2 and polysilicon (poly-Si) interface roughness [10] and random dopant fluctuations (RDF) [11].

Operating environment of the devices on a chip varies as well. Global variations in the supply voltage as well as variations in the local supply grid directly affect the CMOS gate delays, presenting sources of spatially-correlated variability. Operating temperature varies as well, both globally and locally, adding another spatially-correlated component of performance variability.

Device parameters are also variable in time, during the design process or during the chip lifetime. Variations in time include intentional and random changes in the manufacturing process, time-dependent degradation in transistor parameters and changes in supply and temperature. Time-dependent degradation in transistor performance, particularly due to bias temperature instability (BTI), is a major concern. Negative BTI (NBTI) is caused by trapping of the carriers in the PMOS gate interfaces under high biases, which causes threshold increase and degraded current. BTI, which only affected PMOS transistors in $Si\text{-}O_2$ gate stacks, now affects both NMOS and PMOS transistors in high-K metal gate devices [12].

Random telegraph signal (RTS) noise is also another time-dependent source of variability that is becoming a significant concern in design with highly scaled transistors. It is estimated that V_{th} fluctuation due to RTS will exceed V_{th} variation due to RDF at 3 sigma levels at the 22nm technology node [13].

Chip yield is the probability that a chip is both functional and meets the parametric constraints, such as timing and power. A circuit with more design margin will have a higher yield, as it will be more immune to variability. The challenge is in finding the smallest margin necessary for the required yield so that performance is not overly constrained, resulting in large power overhead. The appropriate design margin depends on the type of design, circuit style, its function and use. This paper focuses on the variability impact on combinatorial logic, sequential logic and embedded static random access memory (SRAM), as three distinct digital circuit styles that require different margins.

III. Variability Characterization in Space and Time

In order to incorporate variability in the design, it is necessary to characterize it. Technology variability is characterized during the technology development phase and is monitored during the manufacturing process. Conventional test structures focus on the extraction of the I-V and C-V characteristics of the devices and the interconnect for model corners, while a simple subset of structures is placed in the wafer's scribe lines for continuous process monitoring.

The measured device data is fitted to a compact (SPICE) model and some aspects of variability are captured in the statistics of the model parameters. This information is used to generate process corners and perform Monte Carlo simulations or statistical timing analysis of the circuits. However, it does not consider the spatial correlation of devices and typically does not differentiate between within-die and die-to-die variations. Furthermore, systematic variations due to strain, proximity effects, and time-dependent variations such as BTI and RTS noise are not well modeled and are treated as random. All this leads to overly conservative design margins in advanced processes.

Characterizing more details of variability using suitable test structures allows designers to reduce margins for systematic variations and, with the help of statistical timing and optimization tools, use the right amount of margin to obtain an optimal design that maximizes performance, power and yield.

Device arrays spread over a large chip area with fine spatial resolution provide information on within-die statistics and spatial correlation. Measuring many chips from several wafers and wafer-lots provides die-to-die variability information [14][15]. Tracking the location of the measured devices with respect to the chip, the reticle and the wafer provide a means of locating systematic variation in the manufacturing process, allowing the foundry to correct the variation or, allowing designers to absorb the impact of the variation in the design. Averaging the data for an array of devices during the measurements suppresses random variation and exposes systematic effects.

A. Logic Characterization

The design of logic characterization structures faces several tradeoffs, some of which are illustrated in Figure 1: spatial resolution vs. spatial coverage, the ease of data analysis vs. the sensitivity to the desired parameter or the design complexity vs. the test time.

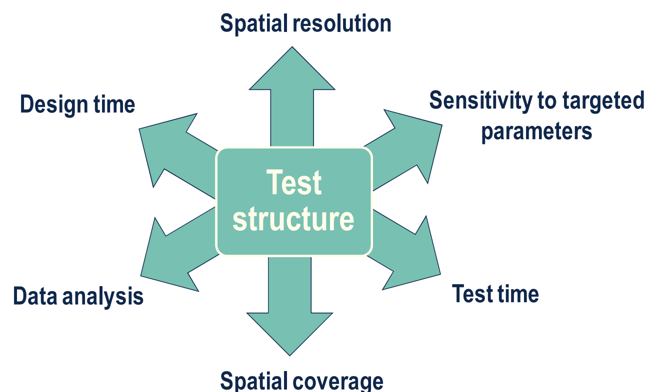

Figure 1: Tradeoffs in test structure design.

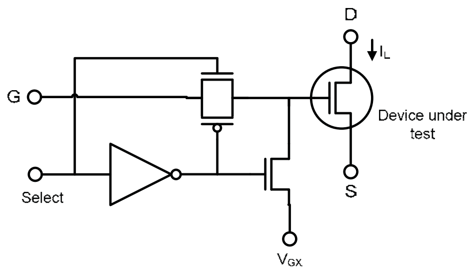

Figure 2: Instance of the device under test in the DMA structure.

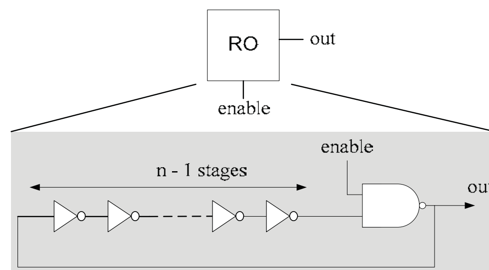

Figure 3: A simple ring-oscillator structure.

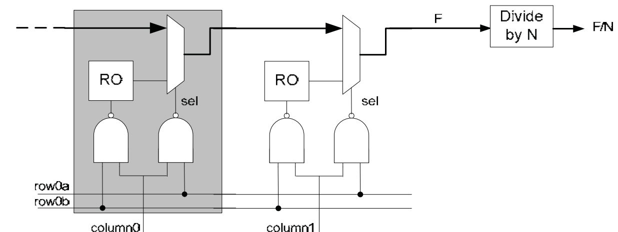

Figure 4: An addressable array of RO. The local RO frequency is divided down and measured off-chip.

Four types of measurement structures are being used: (1): direct measurements of resistances and capacitances [16][17], (2) I-V device measurements [20][21][22] (3) ring oscillators [18][19], and (4) functional blocks [23].

Regardless of the test structure, particular schematic designs and layouts of the devices under test can be biased to explore the impact of processing, such as gate patterning, or circuit topology [4][15][24][25][26].

A common method for characterizing transistor variability consists of measuring the current of individual transistors in an array. Direct measurements of resistances, capacitances and I-V characteristics provide the distributions of desired values, but both the stimuli and the measurements are analog, resulting in long test times and limited datasets.

I-V characteristics of large transistor arrays can be largely collected using device matrix arrays (DMAs), where many transistors can have drains and sources connected to a single bus, while the gate voltages are externally swept, shown in Figure 2 [20][27]. This structure provides a large flexibility in analyzing each device's characteristics and spatial device correlations. The limitation of the structure is in the accuracy of subthreshold current measurements because of many devices that share the same line. An array of transistors can be connected and multiplexed to address the measurement nodes of an individual device under test. Source-drain currents can be measured with reasonable speed and accuracy. Measurement of gate leakage and subthreshold current requires very sensitive measurement instruments to measure small currents or very large devices to generate bigger currents. Parasitic currents need to be removed with calibration and noise can be reduced by averaging the measurement over time. Using larger devices to get more current is a compromise with having a finer spatial resolution.

On the other hand, variability characterization using ring oscillators (ROs) is commonly performed for high characterization speed and simple, frequency measurements [28]. Variation in the frequency of the ROs is related to variation of device parameters that affect transistor switching speed and capacitive load. RO test structures are constructed by using inverting gates such as inverters or NAND/NOR gates, Figure 3. An array of ROs can be addressed and individual RO frequency can be divided down and measured off-chip, Figure 4. Measurement of large arrays, therefore, can be performed in a fast, accurate and automated manner.

Since a RO averages the gate delays, spatial resolution of variability is limited to the spacing between the ROs. This essentially eliminates transistor-level variations and averages random variation over the RO. The longer the RO the less observable is the random component of variation. Nevertheless, this method measures the variability of the switching speeds of a gate with a spatial resolution that is smaller than the logic depth of most datapaths.

Designing gates to be more sensitive to certain process parameter can help to correlate the RO frequency variation with that parameter. Differential measurements of two structures that are identical in all aspects except for a certain layout parameter can be used to isolate particular effects. This has been successfully employed to measure the impact of layout on transistor performance [15][4] and in the measurement of the effects of NBTI [29]. It has been effective in isolating the effects of lithography on the effective channel length, gate proximity, impact on density and diffusion length on strain and STI-induced stress.

Making electrical measurements under different conditions can help to estimate process parameter variation. For example, by measuring the subthreshold current and RO frequency at different temperatures, supply and substrate voltages, the variation of the main process parameters responsible for systematic variations can be extracted. The same technique has been used to reveal the impact of strain and proximity on device properties.

B. Characterization of gate delays

Ring oscillators efficiently characterize systematic variations, but average out random, spatially uncorrelated effects. A simple concept of variable-length ring oscillators [30], can measure the delay of a pair of gates. This is accomplished by measuring the frequency differences between differently configured ring oscillators in Figure 5 [26].

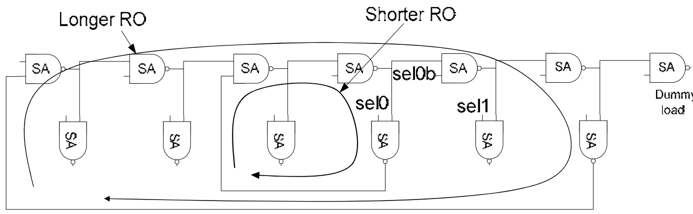

Figure 5: Variable-length ring oscillator.

Ring oscillators can be modified to allow for multiplexing of individual delay elements for delay characterization of individual gates [31].

The idea for characterizing individual delays through path differences has been used for on-chip measurements of flip-flop setup times and clock-to-output delays. Individual direct delay measurements have been performed for flip-flop setup and clock-output delays [32]. By placing appropriately configured flip-flops into ring-oscillator configurations, variations in the setup times and clock-to-output delays can be measured [33].

C. SRAM Characterization

Regular layouts have allowed aggressive scaling of SRAM transistors compared to combinational logic. Simultaneously, these highly scaled devices with design rules relaxed compared to digital logic, tend to exhibit higher sensitivities to systematic effects in addition to increased random variation. These systematic effects have been attributed to temperature non-uniformities during annealing, STI-induced stress, and process-induced cell asymmetry [34]. Increased variability with technology scaling has a large negative impact on SRAM design. SRAM cells use the smallest transistors available, and therefore are susceptible to largest amounts of random variability, while the technology scaling enables integration of twice as many cells in each new process generation. As a result, it is becoming necessary to satisfy the design where the functionality of the cell is guaranteed more than 7 standard deviations away from the mean, while the standard deviations in threshold voltages are increasing. Therefore, SRAM variability characterization over a wide range of process parameters presents a particular challenge. SRAM yield is guaranteed through appropriate design margins against various failure modes. An array fails if any of its cells cannot be written, loses the value during the read, or cannot retain the value in standby.

Traditionally, SRAM design margins have been estimated through SPICE and TCAD simulations for each of the failure modes, and largely depend on the accuracy of models utilized. However, as processes become increasingly complex and harder to control, along with the reduced device geometries, designers can no longer rely on model accuracy to fully capture the random effects in large cache memories. Recently, methods have been developed to characterize SRAM variability by measuring DC read/write/retention margins in small SRAM macros with wired-out storage nodes. In these macros SRAM is commonly characterized by measuring the I-V characteristics of its constituent transistors or by characterizing the static read stability or static writeability of the SRAM cells [35]. This method requires the insertion of large switch networks to access all internal storage nodes without changing the lithographic environment of the cells, Figure 6. As a result, this approach is limited to smaller data volumes that may be unsuitable for failure analysis of large cache memory. Recent large-scale 3D device simulations have demonstrated that even just random dopant fluctuations cause non-Gaussian distributions of transistor threshold voltages in scaled technologies [36], making it difficult to estimate the behavior of cells in the tail of the margin distributions. Conventional padded-out cell characterization techniques fail to characterize many of these effects due to insufficient spatial resolution. Large-scale characterization techniques, involving characterization of SRAM cells in-situ within the array, provide a better estimate of the impact of these systematic effects on SRAM performance. Thus, SRAM designers continue to rely on collecting distributions of bit-line read currents [37] and minimum operating voltage (V_{MIN}) [38][39] to gauge SRAM read stability and writeability in large functional SRAM arrays.

It has been recently demonstrated that read and write margins of a large SRAM array can be measured by bitline current measurements using the setup in Figure 7, in response to wordline, bitline and supply voltage sweeps [34]. Distributions of read and write margins correlate well with the distributions of Vmin during read and write.

Preserving the structure of the SRAM array allows for collection of orders of magnitude more measured variability data in the same chip area.

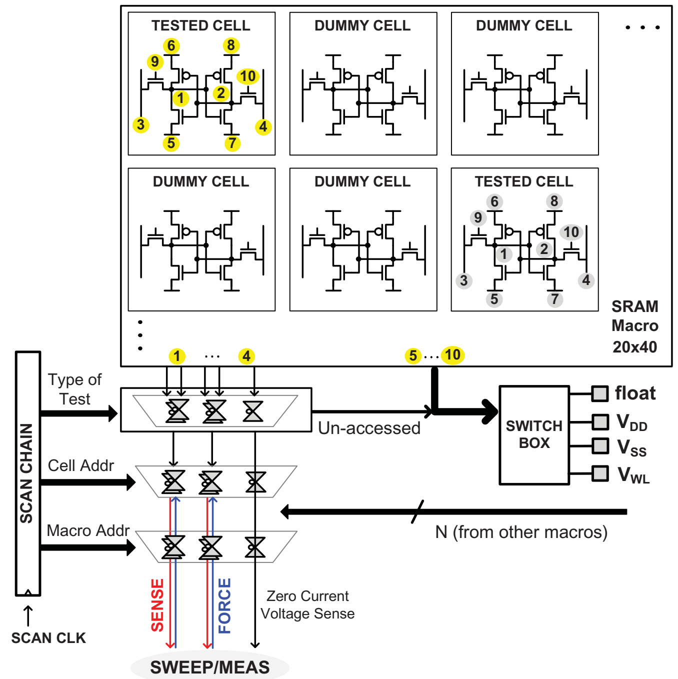

Figure 6: Measurement macro for padded-out SRAM cells.

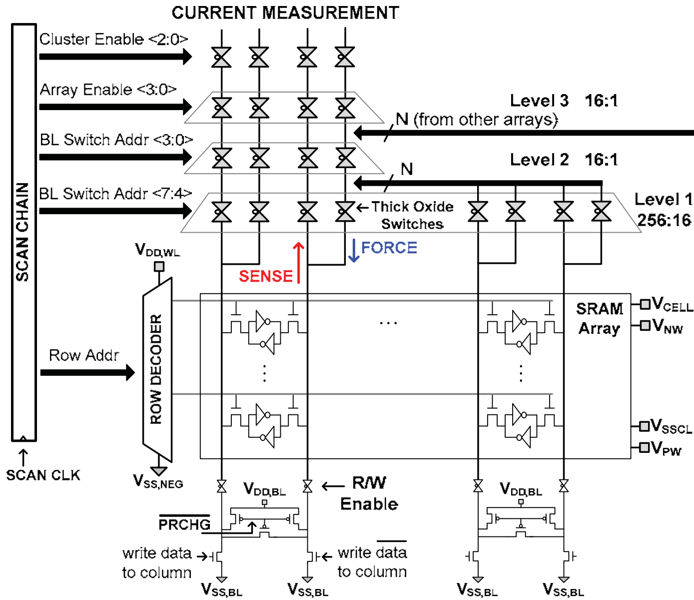

Figure 7: Large-scale SRAM margin measurements [34].

However, static margins only provide a partial picture needed for robust SRAM design. By its definition, static read margin is pessimistic – it assumes that the cell is under the read stress for infinitely long time. Conversely, static write margin is overly optimistic, as it allows infinite time for the cell to be written. In practice, wordlines are pulsed for a short amount of time, during which the cells are written and read. There is a recent trend in assessing dynamic SRAM read/write margins, for more accurate estimation of the necessary operating voltages [40].

Dynamic behavior can be characterized by driving the wordlines with variable pulsewidths, Figure 8. [41]. Shorter pulse widths result in increased read/write failure rates. This method of characterizing the dynamic SRAM behavior is fairly compact and could be embedded in a practical design. Another method for characterizing the dynamic SRAM behavior is through the use of ring oscillators [42][43]. A tunable ring oscillator can be connected to the bitlines; its oscillating frequency would vary with each cell selected through the WL and connected within the RO. An example RO for characterizing write margins is shown in Figure 9.

Repeated measurements using many of the presented structures can be used for characterizing the time-dependent variability of the SRAM characteristics. Simple repetitions of the measurements under the same conditions expose the effects of RTS noise. Measurements under increased supply voltage and elevated temperatures reveal the impact of BTI.

SRAM design in the presence of RTS noise therefore requires accurate characterization of the statistical distributions of V_{th} fluctuation, applied to a statistical model of SRAM failure, in order to budget design margins appropriately. Enhanced characterization techniques have been proposed to speed-up significantly the measurement of these statistical distributions [44].

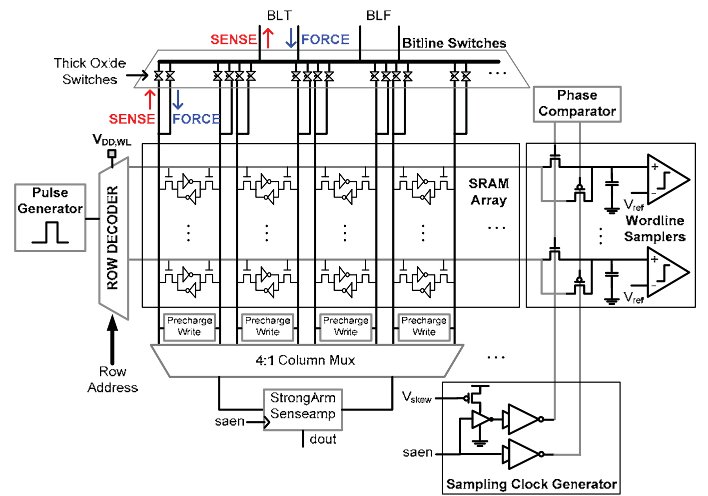

Figure 8: Dynamic SRAM characterization macro [41].

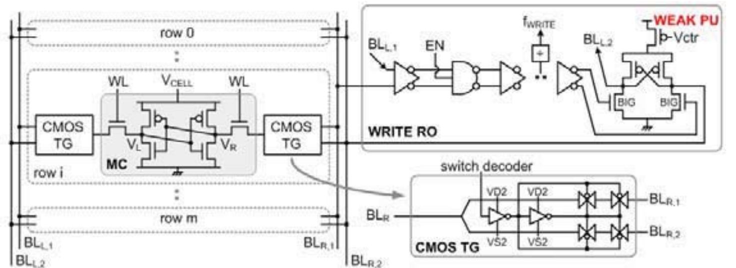

Figure 9: SRAM RO for estimating write margins [43].

IV. Robust SRAM Design

The goal of variability studies is to improve the understanding of the nature of variability, and to help classification into random and systematic components, including spatial and temporal correlations. Accurate modeling of variability will result in a design with smaller design margins for desired yield.

Guaranteeing yield for a large array is a challenging statistical optimization problem, even with Gaussian distributions of each transistor's parameters. This is because of the non-linear dependence of the margins on the transistor parameters. Figure 10 illustrates the results obtained using characterization macros from Figures 6 and 7. All distributions are Gaussian near the center, but deviate significantly in the tails. The key parameter needed for the design σ/μ varies with the definition of the margin, and all three write and all three read margins have different σ/μ values. These differences exist because of different setups for evaluating the margins; all setups are practically equivalent, but none of them may represent the actual failure mode in a chip.

The point of failure for any of the stability criteria can be found by tracing the variables using the steepest gradient method [45][46]. The method can be accelerated using statistical techniques such as importance sampling [47][46][48] and statistical blockade [49].

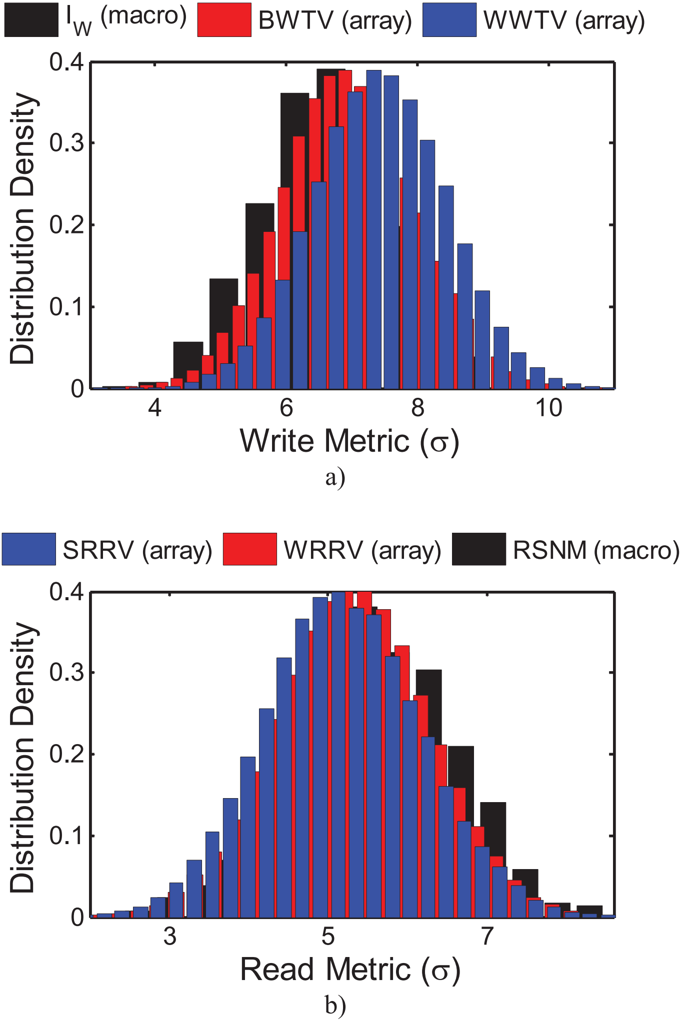

Figure 10: Distributions of various static read and write metrics measured at low voltages: a) Write margins, measured on a padded-out macro (IW) and on an SRAM array (BWTV and WWTW). b) Read margins, measured on a padded-out macro (RSNM) and on an SRAM array (SRRV and WRRV). The metrics are defined in [34].

To improve the read stability or writeability in SRAM, the average margin is increased by adjusting one of the terminal voltages. Lowering the column supply voltage or writing with bitline voltages less than 0V has been used to improve the writeability of the cell, meanwhile, lowering the wordline voltage has been demonstrated to improve the read stability while trading off writeability. While these techniques have been applied at the design time, the design margins can be minimized by tracking the systematic variations in the process. Since the wordline voltage reduction trades off the increased read margin for the reduced write margin, the optimum between the two can be sensed by averaging a number of SRAM cells stressed for both read and write [46].

While all of the current SRAM yield enhancement techniques target systematic components of variability, there is little work in attempting to estimate the tails of distributions. One potentially promising approach is through the use of 'canary' cells – cells that are designed to fail before any of the cells in the array [50].

Time-dependent degradation in transistor performance due to BTI is also a major concern in SRAM. In contrast to logic circuits which typically face alternating inputs, SRAM transistors face the worst BTI conditions when a DC bias is applied to the transistors for a prolonged period of time. The fact that BTI affects both NMOS and PMOS devices in high-k processes makes SRAM V_{DD} margining more complex because $V_{DD,min}$ degrades at different rates depending on whether the original $V_{DD,min}$ distribution was read- or write-margin limited.

RTS noise is a significant concern in SRAM design involving highly scaled transistors, as its magnitude scales faster than the RDF-induced variations. However, experimental results indicate that while large RTS noise magnitude is present in SRAM transistors, the additional margin needed to compensate for RTS is actually much smaller. This is due to the fact that when convolving a long-tailed distribution (RTS) with a normal distribution (RDF), the outliers in the long-tailed distribution have a low probability of being the most probable failure point in the design [44]. Furthermore, both theoretical and experimental analyses suggest that at least some components of RTS noise and BTI stem from the same traps, and therefore should be included in the same margin [51].

V. Variability in Logic

Digital logic utilizes larger devices than SRAM, which reduces both random and some systematic variability components. Digital logic naturally averages random, spatially uncorrelated variations. As a result, longer critical paths reduce the impact of random variability; the σ/μ of random variability decreases with $\sqrt{N}$, where the N is the number of gates in the path. Longest paths in a circuit need to meet the setup time requirement for the receiving flip-flop, which need to be margined appropriately. Shortest paths need to be margined for avoiding the violation of the hold time on the flip-flop. Hold margins are often dictated by the timing mismatches between individual gates, As a result, random variability has less impact on setup time margins than the hold time margins.

Systematic and spatially correlated variations are not averaged and σ/μ is independent of the logic depth.
Designers generally account for systematic and random variability by adjusting the design margins under the statistical static timing analysis (SSTA) [52] and through tracking of critical path delays [53].

The role of the SSTA is to apply appropriate timing margins on the design that account for the systematic variations, and various correlations in the design, to avoid over-constraining.

In a typical VLSI design process, satisfying design corners is deemed necessary and assumed sufficient in order to validate a design. This approach typically regards all variations as D2D, with all devices on a chip having identical

process parameters. WID spatial correlations between the clock and data timing paths present an opportunity for reduction in setup time margin. Measurements in earlier technology nodes revealed spatial correlation radii of approximately 1mm [18][19]. These correlations are caused by systematic processing effects; however, they are not modeled and therefore appear as random, with a certain degree of spatial correlation. These spatial correlations are of the order of the size of a typical digital block, which makes many paths inside a block partially correlated, allowing for some reduction in margins. However, measurements in 45nm technology reveal no spatial correlations at the block level [4]. This is caused by (1) improvements in processing technologies, (2) reduced gate sizing, which increases true random variability, potentially masking the spatially correlated component. Although not confirmed, it is believed that the reticle-level spatial effects are still present [7]. These effects can be used to establish timing correlations for inter-block data and clock distributions in statistical timing analysis.

Spatial, processing-induced correlations are not the only one present in the chip. Layout-induced variations are common for all gates with the same topology or with the same neighborhood, and are therefore systematic. However, since many of the effects are not modeled, they appear to the designer as random. They can be corrected by better processing, accounted for during circuit extraction from the layout and acknowledged in the models, or can be treated statistically in the timing analysis.

Flip-flops are topologically the most complex cells in a standard-cell library. Variability affects their clock-to-output delay, setup and hold times in a partially correlated way, since some of the transistors are shared between these timing paths. The variability of these gates is proportional to the stack height, not unlike complex combinatorial gates.

Efficient energy management often relies on operating chips under dynamically varying supply or threshold voltages. The required supply voltage to maintain the required frequency of operation is maintained by monitoring the supply/substrate voltage controlled set of critical path replicas. These replicas attempt to minimize the additive design margin in this system by tracking the most likely critical path for the particular supply voltage in the given design corner. The mix of critical paths must be sufficient to identify the slowest path at each supply voltage, and is often composed by mixing the NAND2 NAND3, NOR2 gates, inverters, interconnect and pass-gates [54].

Measurements have shown that standard deviation of the delay variability increases with the number of transistors in the transistor stack (i.e. random variability in the delay of a NAND3 is larger than in NAND2) [26]. One interesting direction in future designs is to allocate the margin for random variability dependent on the composition of gates in the critical paths, to reflect this issue.

VI. Conclusions

Variability limits the lowest operating voltage for a technology. This presents a challenge for continued scaling, where one of the major scenarios relies on continued improvements in energy efficiency of multicore processors through voltage scaling. To overcome voltage scaling barriers, variability characterization needs to be extended to enable compact, in-situ energy and performance monitoring of logic and memory blocks. Continued improvement in design techniques, which incorporate mitigation of the effects of variability, in addition to continuous performance monitoring would enable operation of high-volume products at near-threshold supplies.

Acknowledgments

The authors wish to acknowledge the support of the C2S2 Focus Center, one of six research centers funded under the Focus Center Research Program (FCRP), a Semiconductor Research Corporation (SRC) entity. This work was also supported by the SRC's Global Research Collaboration task 1600 and IBM Faculty Partnership Award. STMicroelectronics donated chip fabrication. The authors also would like to thank the contributions of students, faculty and member companies of the Berkeley Wireless Research Center. In particular, we are grateful to Prof. Tsu-Jae King Liu, Prof. Costas Spanos, Prof. Andrew Neureuther, Dr. Yasumasa Tsukamoto, Dr. Andrew Carlson, Dr. Radu Zlatanovici, Changhwan Shin, Kun Qian, Lynn Wang, Jason Tsai, Kenneth Duong and Lauren Jones.

References

[1] K.A. Bowman, S.G. Duvall, J.D. Meindl, "Impact of die-to-die and within-die parameter fluctuations on the maximum clock frequency distribution for gigascale integration," *IEEE Journal of Solid-State Circuits*, vol. 37, no.2, pp.183-190, Feb. 2002.

[2] 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 Journal of Solid-State Circuits*, vol. 37, no. 11, pp. 1396-1402, Nov. 2002.

[3] K. Bernstein, *et al*, "High-performance CMOS variability in the 65-nm regime and beyond," *IBM Journal of Research and Development*, vol. 50, no.4-5, pp.433-449, July/Sept. 2006.

[4] L.-T. Pang, K. Qian, C. Spanos, B. Nikolić, "Measurement and analysis of variability in 45nm strained-Si CMOS technology," *IEEE Journal of Solid-State Circuits*, vol. 44, no. 8, pp. 2233-2244, August 2009.

[5] D. Marković, V. Stojanović, B. Nikolić, M.A. Horowitz, R.W. Brodersen, "Methods for true energy-performance optimization," *IEEE Journal of Solid-State Circuits*, vol. 39, no. 8, pp. 1282-1293, August 2004.

[6] S. Nassif, "Delay variability: sources, impacts and trends," *IEEE Int. Solid-State Circuits Conf., Dig. Of Tech. Papers*, San Francisco, CA, Feb. 2000, pp. 368 – 369.

[7] K. Qian, C. J. Spanos, "A comprehensive model of process variability for statistical timing optimization," in *Design for Manufacturability through Design-Process Integration II, V. K. Singh and M. L. Rieger, Eds., Proceedings of SPIE, Vol. 6925, Bellingham, WA: SPIE -- Society of Photo-Optical Instrumentation Engineers*, 2008, pp. 1G-1-11.

[8] A. B. Kahng, Y. C. Pati, "Subwavelength lithography and its potential impact on design and EDA," in *Proc. Design Automation Conference*, New Orleans, LA, USA, June 1999, pp. 799–804.

[9] P. Oldiges, *et al*, "Modeling line edge roughness effects in sub 100 nanometer gate length devices," *Proc. Int. Conf. on Simulation of Semiconductor Processes and Devicces, SISPAD 2000*. Seattle, WA, Sept. 6-8, 2000. pp. 131-134.

[10] A. Asenov, S. Kaya, J.H. Davies, "Intrinsic threshold voltage fluctuations in decanano MOSFETs due to local oxide thickness variations," *IEEE Trans. on Electron Devices*, vol. 49, no. 1, pp. 112-119, Jan. 2002.

[11] D.J. Frank, Y. Taur, M. Ieong, H.-S.P. Wong, "Monte Carlo modeling of threshold variation due to dopant fluctuations," *1999 Symposium on VLSI Circuits*, *Dig. Tech. Papers.* Kyoto, Japan, June 1999, pp.171-172.

[12] J.C. Lin, A.S. Oates, C.H. Yu, "Time dependent Vcc_{min} degradation of SRAM fabricated with high-k gate dielectrics," *Proc.45th Annual IEEE International Reliability Physics Symposium,* Phoenix, AZ, Apr. 15-19, 2007, pp.439-444.

[13] N. Tega, *et al*, "Increasing threshold voltage variation due to random telegraph noise in FETs as gate lengths scale to 20nm," *2009 Symp. on VLSI Tech*. *Dig Tech. Papers*, Kyoto, Japan, June 2009, pp. 50-51.

[14] H. Onodera, "Variability: Modeling and its impact on design," *IEICE Trans Electron.* Vol.E89-C, no. 3, pp. 342-348, March 2006.

[15] L.-T. Pang, B. Nikolić, "Measurements and analysis of process variability in 90nm CMOS," *IEEE Journal of Solid-State Circuits*, vol.44, no.5, pp. 1655-1663, May 2009.

[16] M. Orshansky, L. Milor, C. Hu, "Characterization of spatial intrafield gate CD variability, its impact on circuit performance, and spatial mask-level correction," *IEEE Trans. Semiconductor Manufacturing*, vol. 17, no. 1 pp. 2-11, Jan 2004.

[17] D. Sylvester, J.C. Chen, C. Hu, "Investigation of interconnect capacitance characterization using charge-based capacitance measurement (CBCM) technique and three-dimensional simulation," *IEEE Journal of Solid-State Circuits*, vol. 33, no. 3, pp. 449-453, March 1998.

[18] K. Gonzalez-Valentin, Extraction of variation sources due to layout practices, M.S. Thesis, Massachusetts Institute of Technology, Cambridge, MA, 2002.

[19] J. S. Panganiban, A ring oscillator based variation test chip, M.S. Thesis, Massachusetts Institute of Technology, Cambridge, MA, 2002.

[20] K. Agarwal, *et al*, "A test structure for characterizing local device mismatches," *2006 Symp. VLSI Circuits Dig. of Tech. Papers*, Honolulu, HI, June 2006, pp.67-68.

[21] L.T. Pang, B. Nikolić, "Impact of layout on 90nm CMOS process parameter fluctuations," 2006 Symposium on VLSI Circuits, Digest of Technical Papers, Honolulu, HI, June, 2006, pp. 84-85.

[22] V. Wang, K.L. Shepard, "On-chip transistor characterisation arrays for variability analysis," Electronics Letters , vol.43, no.15, pp.806-807, July 19 2007.

[23] N.Drego, A. Chandrakasan, D. Boning, "All-digital circuits for measurement of spatial variation in digital circuits," *IEEE Journal of Solid-State Circuits*, vol. 45, no. 3, pp. 640-651, March 2010.

[24] B. Wan, J. Wang, G. Keskin, L. T. Pileggi, "Ring oscillators for single process-parameter monitoring," *IEEE Workshop on Test Structure Design for Variability Characterization*, San Jose, CA, November 2008.

[25] L. T.-N. Wang, L.-T. Pang, A. R. Neureuther, B. Nikolić, "Parameter-specific electronic measurement and analysis of sources of variation using ring oscillators," *Proc. SPIE 7275*, San Jose, CA, February 22-27, 2009, pp. 72750L-7275L-10.

[26] J.-H. Park, L.-T. Pang, K. Duong, B. Nikolić, "Fixed- and variable-length ring oscillators for variability characterization in 45nm CMOS," *IEEE Custom Integrated Circuits Conference*, San Jose, CA, September 13-16, 2009. pp. 519-522.

[27] S. Ohkawa, M. Aoki, H. Masuda, "Analysis and characterization of device variations in an LSI chip using an integrated device matrix array," *IEEE Transactions on Semiconductor Manufacturing*, vol. 17, no. 2, pp.155-165, May 2004.

[28] M. Bhushan, A Gattiker, M.B. Ketchen, K.K. Das "Ring oscillators for CMOS process tuning and variability control," *IEEE Trans. on Semiconductor Manufacturing,* vol. 19, no. 1, pp. 10–18, Feb 2006.

[29] T. Kim, R. Persaud, and C. H. Kim, "Silicon Odometer: An on-chip reliability monitor for measuring frequency degradation of digital circuits," *IEEE Journal of Solid State Circuits*, vol. 43, no. 4, pp. 874–880, Apr. 2008.

[30] B. Zhou, A. Khouas, "Measurement of delay mismatch due to process variations by means of modified ring oscillators," *Proc. 2005 IEEE International Symposium on Circuits and Systems*, ISCAS'2005, Kobe, Japan, May 2005, pp. 5246-5249.

[31] B.P. Das, B. Amrutur, H.S. Jamadagni, N.V. Arvind, V.Visvanathan, "Within-die gate delay variability measurement using re-configurable ring oscillator," Proc. *IEEE Custom Integrated Circuits Conference*, San Jose, CA Sept. 2008.pp 133-136.

[32] N. Nedovic, W.W. Walker, V.G. Oklobdzija, "A test circuit for measurement of clocked storage element characteristics," *IEEE Journal of Solid-State Circuits*, vol.39 , no. 8, pp. 1294-1304, Aug. 2004.

[33] K. Duong, Circuits for Measurement of Flip-Flop Performance Variability, M.S. Thesis, University of California, Berkeley, 2008.

[34] Z. Guo, A. Carlson, L.-T. Pang, K. Duong, T.-J. King Liu, B. Nikolić, "Large-scale SRAM variability characterization in 45nm CMOS," *IEEE Journal of Solid-State Circuits*, vol. 44, no.11, pp. 3174-3192, Nov. 2009.

[35] A. Bhavnagarwala, *et al*., "Fluctuation limits and scaling opportunities for CMOS SRAM cells," *IEEE International Electron Devices Meeting, IEDM, Tech. Dig*., 2005, pp. 675-678.

[36] D. Reid, C. Millar, G. Roy, S. Roy, A. Asenov, "Analysis of threshold voltage distribution due to random dopants: A 100 000-sample 3-D simulation study," *IEEE Transactions on Electron Devices*, vol. 56, no. 10, pp. 2255-2263, 2009.

[37] T. Fischer, *et al*., "Analysis of read current and write trip voltage variability from a 1-MB SRAM test structure," *IEEE Transactions on Semiconductor Manufacturing*, vol. 21, no. 4, pp. 534-541. Nov. 2008.

[38] M. Agostinelli, *et al*., "Erratic fluctuations of SRAM cache VMIN at the 90nm process technology node," *IEEE International Electron Devices Meeting*, *IEDM, Tech. Dig*., 2005, pp. 655-658.

[39] M. Ball, *et al*., "A screening methodology for V_{MIN} drift in SRAM arrays with applications to sub-65nm nodes," *IEEE International Electron Devices Meeting, IEDM'2006, Tech. Dig*., 2006, pp. 1-4.

[40] D. Khalil, et al, "Accurate estimation of SRAM dynamic stability," *IEEE Transactions on VLSI*, vol. 16, no. 12, pp. 1639-1647, Dec. 2008.

[41] S.O. Toh, Z. Guo, B. Nikolić, "Dynamic SRAM stability characterization in 45nm CMOS," *2010 Symposium on VLSI Circuits*,*Dig. Tech Papers.* Honolulu, HI, June 16-18, 2010.

[42] G. D. Carpenter, *et al*, "Pulsed ring oscillator circuit for storage cell read timing evaluation," U.S. Patent No. 7,409,305, Aug. 5, 2008.

[43] J. Tsai, S.O. Toh, Z. Guo, L.-T.Pang, T.-J. King, B. Nikolić, "SRAM variability characterization using tunable ring oscillators in 45nm CMOS," *2010 IEEE International Solid-State Circuits Conference*, *Dig. Tech Papers.* San Francisco, CA, February 7-10, 2010. pp. 354-355.

[44] S.O. Toh, *et al*, "Impact of random telegraph signals on Vmin in 45nm SRAM," *IEEE International Electron Devices Meeting IEDM*, *Tech Dig.* Dec. 2009. pp.768-770.

[45] Y. Tsukamoto, *et al*, "Worst-case analysis to obtain stable read/write DC margin of high density 6T-SRAM-array with local Vth variability," *Proc. IEEE International Conference on Computer-Aided Design, ICCAD'05*, San Jose, CA, 2005. pp.398-405.

[46] A.E. Carlson, *Device and Circuit Techniques for Reducing Variation in Nanoscale SRAM*, Ph.D. Dissertation, UC Berkeley, 2008.

[47] R. Kanj, R. Joshi and S. Nassif, "Mixture importance sampling and its application to the analysis of SRAM designs in the presence of rare failure events," *Proc. Design Automation Conference,* DAC, 2006, pp. 69 – 72.

[48] L. Dolecek, M. Qazi, D. Shah, A. Chandrakasan, "Breaking the simulation barrier: SRAM evaluation through norm minimization," *Proc. IEEE International Conference on Computer-Aided Design, ICCAD 2008*,San Jose, CA, 2008, pp. 322-329.

[49] A. Singhee, R. Rutenbar, "Statistical blockade: a novel method for very fast Monte Carlo simulation of rare circuit events, and its application," *Proc. DATE 2007*, pp. 1–6.

[50] J. Wang, B.H. Calhoun, "Canary replica feedback for near-DRV standby VDD scaling in a 90nm SRAM," *IEEE Custom Integrated Circuits Conference, CICC'2007,* San Jose, CA, Sept. 2007.

[51] Y. Tsukamoto, *et al*, "Analysis of the relationship between random telegraph signal and negative bias temperature instability," *2010 IEEE International Reliability Physics Symposium*, Anaheim, CA, May 2-6, 2010.

[52] C. Visweswariah, *et al*, "First-order incremental block-based statistical timing analysis," *IEEE Transactions on CAD*, vol. 25, no 10. Pp. 2170-2180, Oct. 2006.

[53] T.D.Burd, T. Pering, A. Stratakos, R. Brodersen, "A dynamic voltage scaled microprocessor system," *IEEE International Solid-State Circuits Conference, ISSCC Dig. Tech. Papers*, Feb. 2000. pp. 294-295.

[54] A. Drake, *et al*, "A distributed critical-path timing monitor for a 65nm high-performance microprocessor, *IEEE International Solid-State Circuits Conference, ISSCC'2007*, San Francisco, CA, Feb. 2007. pp. 398-399.