Variability Aware Modeling for Yield Enhancement of SRAM and Logic

Miguel Miranda, Paul Zuber, Petr Dobrovolný, Philippe Roussel
CMOS Technology Department, Process Technology Division, imec, Belgium

***Abstract*—Anticipating silicon response in the presence or process variability is essential to avoid costly silicon re-spins. EDA industry is trying to provide the right set of tools to designers for statistical characterization of SRAM and logic. Yet design teams (also in foundries) are still using classical corner based characterization approaches. On the one hand the EDA industry fails to meet the demands on the appropriate functionality of the tools. On the other hand, design teams are not yet fully aware of the trade-offs involved when designing under extreme process variability. This paper summarizes the challenges for statistical characterization of SRAM and logic. It describes the key features of a set of prototype tools addressing that required functionality together with their application to a number of case studies aiming at enhancing yield at product level.**

I. Introduction

Advances in CMOS VLSI circuit design has primarily relied on technology improvements derived from technology scaling. However, process variability and especially its intra-die uncorrelated portion has significantly increased the uncertainty in the response of sub-45nm CMOS circuits. Today, intra-die (also known as local random) process variations are main contributors to statistical circuit responses (e.g., delay and power). Such increase of process variability at every technology node has imposed new challenges to the design and characterization of SRAM and logic cells.

By placing a constraint either on a system level metric (like cycle time, power) in case of logic and/or on stability metrics (as pass/fail checkpoints) in case of SRAM, we focus on the yield loss component caused by parametric deviations in the electrical device parameters. Defect related yield is not addressed in this paper. Its treatment requires well known separate analysis techniques [1] and it can be considered orthogonal to parametric and functional yield analysis. This way the most likely reasons for statistical failure can be anticipated at design time so as to correct weak design spots before tape-out, hence avoiding costly silicon spin iterations.

For logic, approaches based on sensitivity analysis for statistical library characterization have been adopted by several EDA companies. These are essential tools to feed Statistical Static Timing Analysis (SSTA) [2], [3] and other parametric yield prediction flows. They are simple to implement by assuming linear dependencies between the process parameter variations and the timing related metrics of the cells. Yet they usually ignore the underlying correlation among device variation parameters, hence many times leading to inaccurate response prediction.

As a consequence, these sensitivity analysis techniques have not been yet widely adopted by the design industry because of their lack of accuracy.

In the lack of statistical characterization tools providing sufficient accuracy, CPU time expensive Monte Carlo (MC) loops embedded on existing circuit simulators [4] are the most pragmatic alternative for the characterization of standard cell libraries. However such approaches require a large number of simulation runs. Indeed, thousands of simulations are needed for accurately capturing the tails of the distribution of the affected metrics, typically at a 3σ distance from the average value, hence discouraging designers from their use and becoming a showstopper for the wide adoption of SSTA and other yield prediction analysis flows. Approaches achieving near MC accuracy with a speedup improvements of orders of magnitude are therefore urgently demanded for efficient statistical library characterization.

For SRAM the situation is even worse. Virtually no commercially available solution exists yet. Several issues make memories especially challenging:

1 Usually the nominal simulation of a full memory is reduced to the access path netlist, assuming every other path behaves the same. However this approach particularly fails under local process variations where device to device uncorrelated variability makes every bitcell access operation to behave differently. Since a memory is as good as its worst path, the memory statistics for instance of the read margin or access time is the distribution of the worst of all its cells. As a result, simulating the access path netlist under variability does not model the full memory statistics correctly.

2 Approaches considering the bitcell and sense amplifier together only without its periphery [5] manage to reduce the sample sizes and transistor counts effectively, but also entails incomplete analysis. We exemplify this in Figure 1, which shows how the different components influencing the read operation of the cell can affect its read margin. Indeed, variations can affect the sense amplifier offset and the timing circuit that controls its activation, the row decoder that enables the word line activation, and especially the cell's capability to discharge the bitline. Accounting for the worst case situation of each of these effects would lead to pessimistic estimations of the read margin's probability. As a consequence, the entire equivalent circuit's operation must be simulated under variability to obtain realistic results.

978-3-9810801-7-9/DATE11/©2011 EDAA

3 On top of that, attention must be paid to architectural correlations of bitcells and sense amplifiers, row drivers and other memory parts. A worst case cell instance is not necessarily in the same path with the worst case sense amplifier or the worst case row driver logic so that a blind worst case combination would lead to over-pessimistic results again.

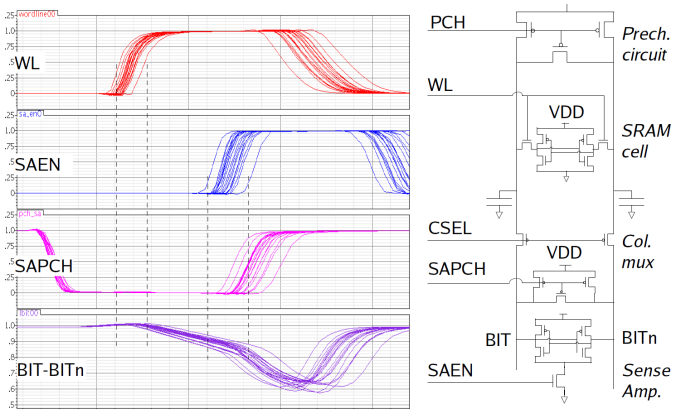

Fig. 1. Read voltage variability (bitline voltage difference at sense amplifier activation time) originates from variations in the timing circuit (precharge and sense amplifier enable times), the bitcell drive strength, the wordline and driver, sense amplifier, and column multiplexers.

In this paper we summarize a set of existing state-fo-the-art automated techniques addressing the required functionality for logic and SRAM. For logic we show how preserving the underlying correlation among process variation parameters leads to overcoming the accuracy limitations of sensitivity based analysis approaches. For SRAM, we present several case studies showing how to capture non-trivial statistical interactions between the cells and the periphery, which remain uncovered when only using statistical electrical simulations of the access path or applying a corner analysis approach.

The automated techniques provide the designer with valuable information on what performance metrics to expect after manufacturing. Since this feedback takes place at design time, a significant reduction in development time and costly silicon iterations is achieved.

II. Variability Aware Modeling

Our flow starts with extracting information about process variability from statistical device compact models and/or silicon measurement data. It uses a complementary set of variability scaling rules such as area dependent mismatch model [6] for scaling local random variations from a reference device size to the actual size as found in the transistor level netlist. In this way we can model most of the known variability effects, either systematic or random. Instead of considering the process variations parameters present in the given statistical device model we prefer to *extract* a set of compact model card independent parameters in the form of device variations in Vt and β and *inject* them in the circuit netlist. This is done by inserting one voltage source in series with the transistor gate (modeling ΔVt) and one current source in parallel with the transistor (modeling $\Delta\beta/\beta$, where β is related to the source-drain current gain).

At this stage, we simulate the circuit for selected combinations of ΔVt and $\Delta\beta/\beta$ variations, and based on its response, we build a regression model of the circuit that predicts its response to any other combination of the variation parameters without involving the use of massive analog simulations. Circuits under consideration range from standard cell descriptions to SRAM access path netlist.

A. Statistical Circuit Response Prediction

For transistor level circuits the challenge is to achieve near MC accuracy with a very limited set of simulations. This is equivalent to building a regression model of the circuit that could predict its response to any change on its input variation parameters. This is achieved by our Variability Aware Statistical Design of Experiments approach, hereafter called s-DoE. Using s-DoE, about two orders of magnitude of better accuracy in the tail response can be achieved compared to sensitivity analysis. Yet it features an electrical simulation effort linear to the cell complexity, hence comparable to sensitivity analysis and being about 10-100 times less CPU time intensive than MC.

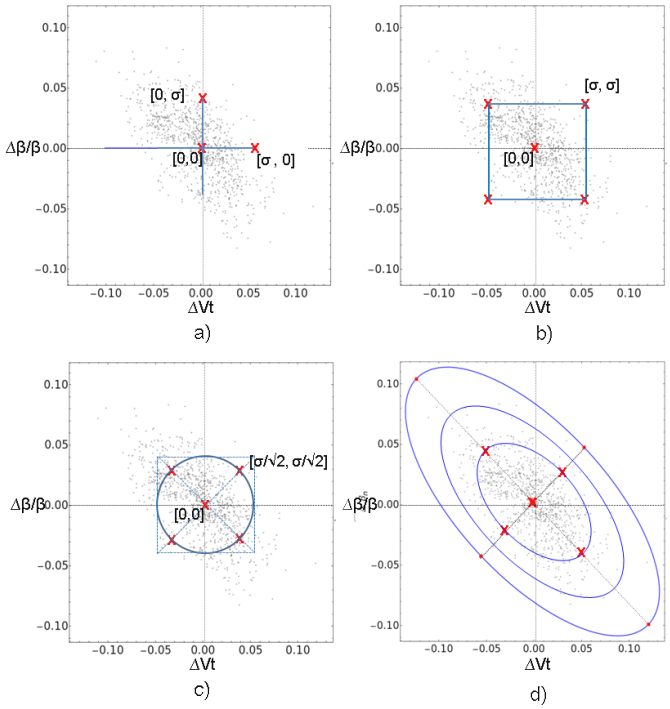

Fig. 2. Possible DoE selection strategies for process variability modeling: a) DoE selection for sensitivity analysis; b) based on full factorial design, c) improvement of full factorial design to avoid pessimistic corners, and d) s-DoE

The concept behind the s-DoE gets illustrated in Figure 2, also in comparison to existing DoE alternatives, from the most simple to more elaborated ones. Sensitivity analysis (see Figure 2.a) is a very simple yet powerful form of DoE where points are positioned along the parameters axes. The drawback

is that the selected points ignore the existing correlations between variation parameters. Therefore it becomes impossible to build response models predicting cross term interactions between them.

Full factorial design (see Figure 2.b) is an improvement over sensitivity analysis since the existing correlations between parameters (p_1 and p_2) are now captured by the selected DoE points. Still full factorial design may lead to selected DoE points which lie outside the area of relevance of the statistical parameter domain (e.g., see the upper right and lower left points of the box in the Figure) which in statistical terms not as relevant. Improvements over full factorial design exist to eliminate over-pessimism (see Figure 2.c), but still the risk of selecting statistically non-relevant points remains.

The s-DoE detects the statistical correlations present in the parameter domain and places the DoE points along their main correlation axes so that all selected points have an equal weight regarding heir representativeness for the statistical population (see inner points in Figure 2.d). In addition, the s-DoE also allows for defining the scope of validity of the fitted response model at any given *distance* from the nominal point. Such distance is representative area of the distribution in which the model is intended to provide accuracy (see outer points in Figure 2.d).

Based on such a selection of representative design points, the method runs electrical simulations on these and based on their response then builds a non-linear regression model to represent the statistical circuit response. This is done in two steps:

1 The first step consists on performing a careful Design of Experiments (DoE) selection. The goal of this stage is to find N_{doe} points which are representative for the n-dimensional input space. The points need to be selected in such a way that they cover as much as possible the volume contained within the statistical input parameter distribution. This is done by first building an n-dimensional PDF representing the multivariate statistic followed by identifying the appropriate set of $2n+1$ minimum amount of s-DoE points (being n the number of variation parameters, hence $2 \times no.devices$). Moreover, the selected DoE should be positioned in the parameter space in such way that they capture the existing correlations among all statistical variables of such PDF. Figure 3 illustrates the selected s-DoE points for an inverter illustrating the strong correlation between the ΔVt and $\Delta\beta/\beta$ variation parameters for an n-FET device in a 32nm technology node.

2 After selecting the s-DoE points and performing an electrical circuit simulation on them, the second step derives a surface model predicting all remaining statistical responses. For that purpose, we use an algorithm for searching the space of possible approximations and, without manual intervention or any previous knowledge about the circuit response (delay, power, etc) it provides the best non-linear function to approximate that response.

In this paper, we focus on use cases and quantitative results for industrial circuits and we only provide a summary of our s-DoE selection approach. Details are left out for a follow-up publication.

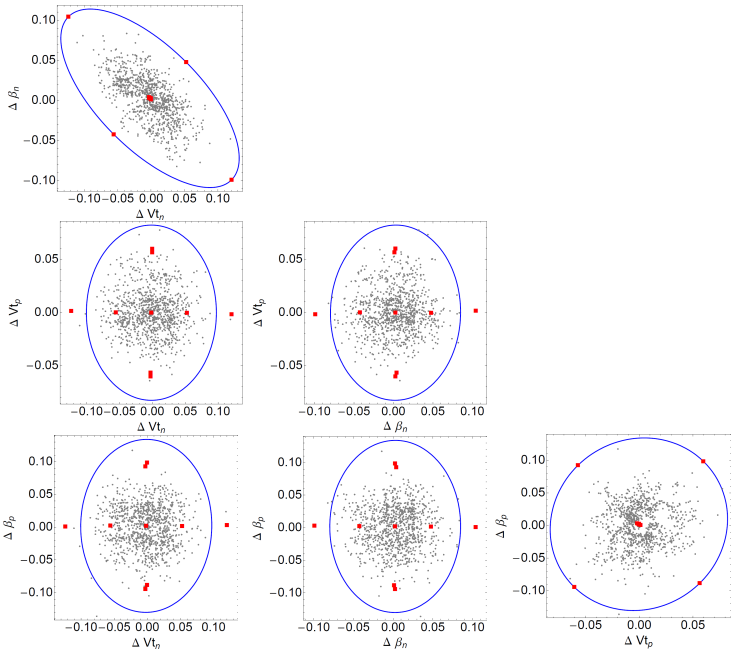

Fig. 3. Pair wise 2-D scatter plots of all possible δV_t and $\delta\beta$ combinations for the two MOSFET of an inverter. s-DoE points are the red dots

B. Statistical SRAM Analysis

For SRAM, the challenge is to capture all (non-trivial) *memory-wide* statistical interactions between the SRAM cell and the periphery, not addressed when using statistical electrical simulations of the access path alone.

For that purpose we have developed a method for statistical memory analysis [8] that relies on a mix of sensitivity analysis of the memory access path to variability in 'islands'. An 'island' is formally defined as a unique and exclusive set of transistor of the path. Islands, such as the timing block, bitcells, IO blocks or word lines, are typically instantiated once in the 1 path netlist, but differ in the number of instantiations in a full memory.

The method of quantitatively predicting memory performance and yield under transistor variability comprises two main steps (see Figure 4:

1 Characterizing Memory Islands (step 1): to derive the sensitivities of one path to variations in certain memory islands, by injecting variability into the island transistors and simulating the modified memory path netlists as described in Section II. We record the resulting sensitivity populations, one for each island. During this step, statistically correlated parametric data and pass/fail information obtained from the access path simulations is collected via inserted measurement and check point statements in the netlist.

2 Architecture Aware Scaling (step 2): to recover the full memory statistics from the island statistics and a specification of the topology of these islands. It populates the

statistics of all paths of the memory by combining the results of 1), under awareness of the memory architecture (topology, organization and redundancy mechanisms), and selects the worst path to represent a memory observation. To build the memory population, also this step comprises a (plain) MC loop. This way, statistical information on the access path percolates to the complete SRAM organization level, resulting in a realistic prediction of the yield as perceived by the memory tester and/or equivalent BIST (built-in-self-testing) technique.

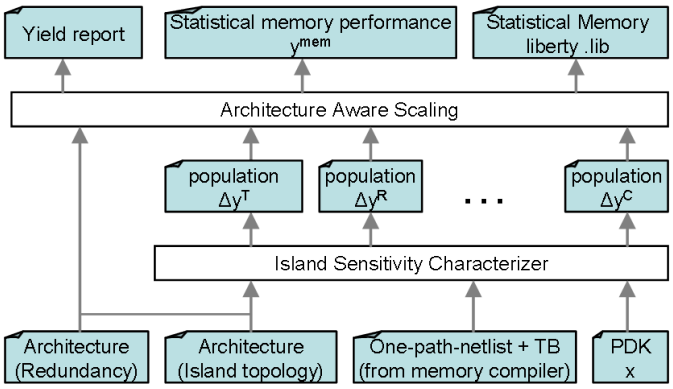

Fig. 4. Overview of the MemoryVAM approach

Key in this strategy is the ability to complement the analysis of a nominal memory model under test with statistically sampled variants of the devices. For that purpose the use of either classical statistical sampling techniques (e.g., importance sampling [9], [10]), or more novel ones such as our Exponent MC enhancement [11]. Also, the most recent regression estimation based techniques (e.g., s-DoE II-A) are best used to significantly reduce the number of statistical simulations needed to achieve a particular level of confidence.

III. Related Work

For logic, the use of RSM techniques in VLSI design for standard cell characterization is not new and its use was originally proposed in the late 80's by [12], [13]. Recently, the use of these regression modeling techniques raised interest again as an effective technique to cope with the explosion on the required process corners to capture the combined impact of local and global process variations [14].

However, all these works are based use of conventional DoE methods like Central-Composite-Design, full factorial and/or Box-Behenken Design [7] that do not consider the statistical nature of the underlying process variation parameters, hence fail to capture their statistical correlations, especially at the tails of the distributions as illustrated in Figure 2.d. Using a minimum of $2n + 1$ points we guarantee a model with cross-terms providing much better accuracy than the arbitrary selection of points used in the majority of these works.

Indeed, unlike conventional DoE approaches, s-DoE selects only design points that are statistically relevant to the parameter domain distribution (see Section I, Figure 2.d). By properly capturing the existing correlation between input parameters, it allows the simulator model to *carry* their effect to the outputs, otherwise leading to inaccuracies in the statistical properties of response model that is build upon these outputs.

For SRAM, the problem of verifying the interactions between the cell and all possible combinations of connecting blocks along all the paths in a memory has not yet been properly addressed in the literature. Chen et al. [15] pointed out the influence of row driver and sense amplifier variability on certain stability metrics and provided a hardware detection of new fault types. Failure mechanisms of the (isolated) cell under local random process fluctuations have been studied extensively [16], [9], [17], [18], [19], [20]. Aitken et al. applied a branch of Gumbel's extreme value theory to derive estimates for the variability related yield of SRAMs [21] and to place proper margins. It is the only known prior work that discovers and describes the PDF shift towards worse values analytically.

We have combined the considerations above and reported a method [22] and its implementation in a prototype tool hereafter called Memory Variability Aware Modeling - or MemoryVAM in short. It is based on a technique that predicts the correct memory wide statistics of any parameter that can be measured in a SPICE/SPECTRE test bench, such as access time, power, stability checks such as read voltage, and so on.

IV. Application, Results and Benchmarking

This section describes different applications of our Variability Aware Modeling flow to industrially relevant logic and SRAM vehicles and describe few case studies in which the approach is used to improve the yield of the product at the design time, hence before manufacturing.

A. Standard Cell Statistical Characterization

We use a subset of cells from a production level 32nm standard cell library and statistical compact model card. The library generated is compatible with Cadence *Liberty* library format.

For performing the most comprehensive benchmark we performed statistical library characterization by three means:

1) *Monte Carlo:* reference method based on 1000 SPICE simulations;
2) *Sensitivity Analysis:* using a commercially available statistical library characterization tool which requires n+1 runs and performs sensitivity analysis;
3) *s-DoE:* response model approach requiring only 2n+1 SPICE simulations. Being n the total number of variation parameters of the circuit.

We use the statistical extension of Encounter Library Characterizer (ELC) from Cadence as the commercial tool to set up an experimental framework for benchmarking sensitivity analysis against s-DoE. The statistical inputs for ELC are the standard deviations of the variation parameters (σ_{Vti} and $\sigma_{\beta i}$ of each transistor i). The tool then gives as outputs: the nominal simulation $\overline{s}$; and the sensitivities of each response (i.e., delay, transition time) to each input variation parameter (e.g. s_{Vti}, $s_{\beta i}$ of each transistor). Sensitivity analysis is based on

the assumption that the statistical circuit response to variations in the process parameters is approximately by a linear superposition of input sensitivities. Thus, if the variation parameters follow a Normal distribution, the output will also follow a Normal distribution with mean and variance given by:

$$\begin{cases} \mu \approx \overline{s} \\ \sigma^2 \approx \sum_{i=1}^{n} \left[(s_{Vti}\sigma_{Vti})^2 + (s_{\beta i}\sigma_{\beta i})^2 \right] \end{cases} \quad (1)$$

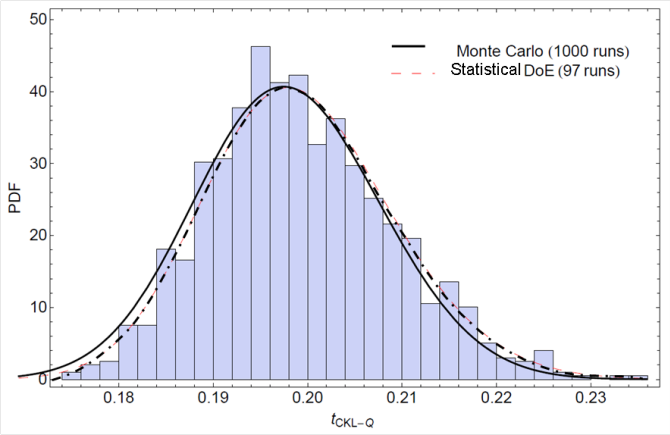

Fig. 5. Histogram showing the distribution of DFFQ Clock-Q delay comparing true response computed using 1000 MC electrical simulations against the 97 $(2 \times (2 \times 24) + 1)$ s-DoE points.

1) Benchmark against Monte Carlo: Figure 5 presents the good agreement observed in the frequency distribution of Clk-Q delay of the flip-flop cell (DFFQ, one of the most complex library cells). This is obtained using s-DoE analysis against MC simulations. Also, Figure 6 plots the agreement observed in the individual response of the simulated points againts those predicted by the regression model for a NAND2 gate. Circles corresponds to the results from 1000 MC HSPICE simulations. Triangles corresponds to the results from 1000 MC responses obtained from the RSM model built using only 17 s-DoE $(2 \times (2 \times 4) + 1)$ HSPICE simulations.

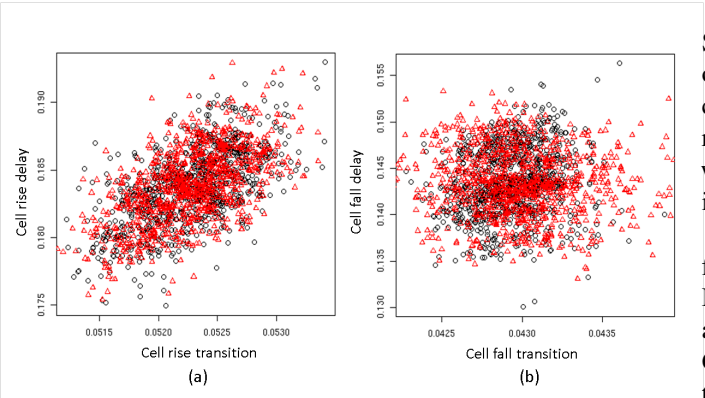

Fig. 6. Comparison of the good agreement of the statistical scatter plots corresponding to a cell delay versus transition time correlation for a NAND2: (a) rise edge; (b) fall edge

2) Benchmark against Sensitivity Analysis: When comparing to Sensitivity Analysis, we are interested on the accuracy of the models when predicting particular circuit responses and not only their statistical properties, especially at the tails of the distribution, which is critical for accurate yield analysis.

One of the advantages of the s-DoE is the ability to choose the region where the model must have high accuracy. The user-defined boundary of interest for the positioning of the s-DoE points allows the designer to focus on the selected region of interest. The consequence is that the error of s-DoE remain small even in the tails of the distribution.

Figure 7 shows the errors of the sensitivity model and s-DoE at the tails of the distributions of hold time and setup time of a DFFQ. These points are distributed at a 3σ distance from the mean. For setup time, ELC shows errors as high as 400%, while s-DoE presents a maximum error of 5%. For the case of hold time violations, commercial tool has errors of 280%, while s-DoE has maximum error of 50%. This proves the ability of s-DoE to accurately predict parametric yield.

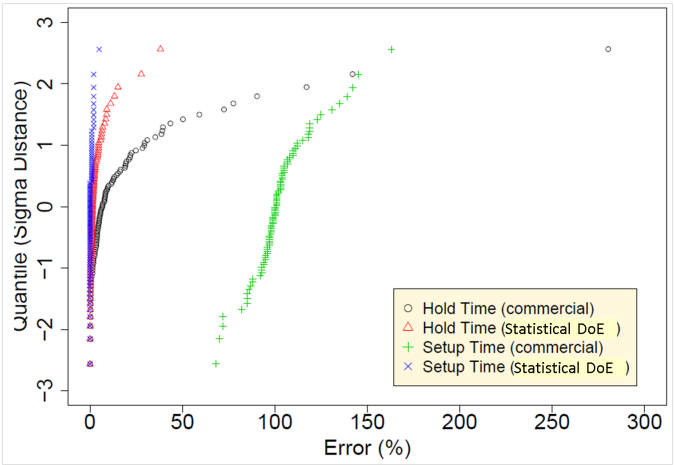

Fig. 7. Error of hold time and setup time of a DFFQ at the tails of the distribution.

B. SRAM Application Case Studies

1) Critical Voltage Analysis in View of Yield: Global Voltage Scaling (GVS) is a useful technique for a dynamic reduction of the memory voltage for minimum power under a timing constraint. Designers use a critical path replica in silicon to report near-failure warnings of all paths, forming a closed loop with the voltage regulator. For GVS one is initially interested in predicting the global variability.

This technique cannot adjust local variability, which it therefore models as a margin. The amount of margin is essential. Is it too small, the risk of timing failures increases. Too high a margin, and the efforts of GVS don't pay off. Again, the GVS engineer must consider the nominal shift of timing due to the many parallel paths existing in the memory. Also the increased timing spread at lower voltages, and the minimum tolerable read-margin require a carefully selected lower bound for VDD. After analysis, our statistical memory analysis tool reveals in Figure 8 that it becomes prohibitive to go below 0.72V. The Sense Amp offset requirements needed to maintain a reasonable yield (¿ 50%) would simply become prohibitive

(e.g., below 30mV). Also that to maintain the same 95% yield when lowering the power supply from 1.16V to 1.02V it is required a Sense Amp with an offset no bigger than 68mV.

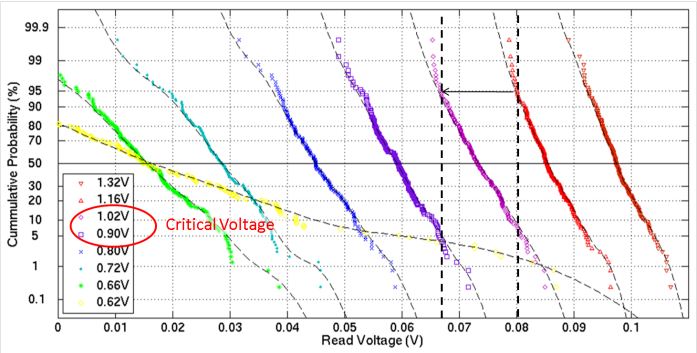

Fig. 8. To maintain yield the Sense Amplifier offset requirements need to become tighter as voltage decreases

2) Yield recovery using circuit knobs: As shown in Figure 1 there is a strong influence of the sense-amplifier activation time to the read-voltage. It has become common in advanced memory design to implement test-time knobs to trade memory speed for robustness in terms of read-voltage. This is done for example with a programmable delay in the sense-amplifier activation signal, which allows more read-voltage to develop. Of course, this artificial delay is subject to variability itself. We have performed an statistical analysis on one of our industry grade memories with this margin control knob in the two extremes of four settings (MCK=00, MCK=11) and compared it to a nave approach that simulates the one-path netlist using regular Importance-sampling. Figure 9 shows the results. As expected, with the aggressive setting, the read-voltage decreases.

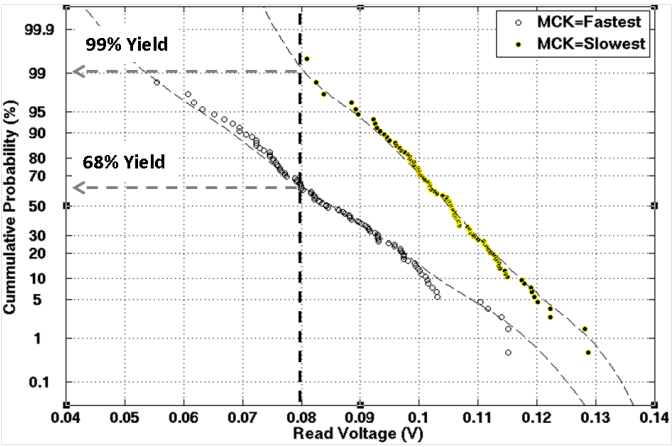

Fig. 9. Setting the margin control knob MCK to 11 causes later sense-amplifier activation time and thus increased read-voltage

One can observe another less intuitive effect. While the nominal point of the read-voltage improves from approximately 170 to 250mV, the median of the memory's read-voltage improves from about 80 to only 105mV. This difference (+70mV vs. +25mV) must result from the fact that the knob is less effective for increasing small read-margins than for increasing high read-margins. There is still a noticeable effect of about 25mV of this knob in the memory, albeit less than predicted by nominal and even traditional statistical SPICE simulation. Of course, this comes at the cost of extra delay, one nanosecond in this case and it also entails a significant short-circuit power overhead. Thus the goal is an economic as possible use of this knob and accurate statistical quantification of its response becomes mandatory.

V. Conclusions

We have summarized two methods featuring accuracy in predicting parametric yield for logic and SRAM circuits and presented their results from benchmarking against standard characterization flows on a set of application case studies.

VI. Acknowledgements

The authors acknowledge the contribution of Lucas Brusamarello on the experiments carried out in this work.

References

[1] R. Ott et al., in ASMC/SEMI: Proc. Advanced Semiconductor Manufacturing Conf. ACM, 1999, pp. 87–91.
[2] C. Visweswariah, et al., "First-order incremental block-based statistical timing analysis," Proc. DAC, 2004. pp. 331–336.
[3] M. Imai, et al., "Non-parametric statistical static timing analysis: an ssta framework for arbitrary distribution," Proc. DAC. 2008, pp. 698–701.
[4] J. G. Amar, "The monte carlo method in science and engineering," *Computing in Science and Engineering*, vol. 8, no. 2, pp. 9–19, 2006.
[5] Y. Zhou, et al. "The impact of beol lithography effects on the sram cell performance and yield" Proc. ISQED 2009, pp. 607–612.
[6] M. Pelgrom et al., "Matching properties of mos transistors," *IEEE Journal of Solid-State Circuits*, vol. 24, no. 5, pp. 1433–1439, Oct 1989.
[7] R. H. Myers, et al. *Response Surface Methodology: Process and Product in Optimization Using Designed Experiments*, 2nd ed. 2002.
[8] P. Zuber, P. Dobrovolny, and M. Miranda, "A holistic approach for statistical analysis of SRAM" Proc. DAC, 2010, pp.717-724.
[9] R. Kanj, et al., "Mixture importance sampling and its application to the analysis of sram designs in the presence of rare failure events," Proc. DAC, 2006, pp. 69–72.
[10] D. Hocevar, et al., "A study of variance reduction techniques for estimating circuit yields," Trans. CAD , vol. 2, no. 3, pp. 180–192, July 1983.
[11] P. Zuber, et al, "Using exponent monte carlo for quick statistical circuit simulation" Proc. PATMOS, 2009.
[12] A. Alvarez, et al., "Application of statistical design and response surface methods to computer-aided vlsi device design," *IEEE Trans. on CAD*, vol. 7, no. 2, pp. 272–288, Feb 1988.
[13] D. Hocevar, et al. , "Parametric yield optimization for mos circuit blocks," *IEEE Trans. on CAD*, vol. 7, no. 6, pp. 645–658, Jun 1988.
[14] J. Kim, et al., "Fast, non-monte-carlo estimation of transient performance variation due to device mismatch," Proc DAC. 2007, pp. 440–443.
[15] Q. Chen et al. "Modeling and testing of SRAM for new failure mechanisms due to process variations in nanoscale CMOS" Proc. VLSI Test Symp., 2005, pp. 292-297
[16] K. Agarwal et al. "Statistical analysis of SRAM cell stability" Proc. DAC , 2006, pp. 57-62
[17] J. Wang et al. "SRAM parametric failure analysis" Proc. DAC, 2009, pp.496-
[18] S. Mukhopadhyay et al. "Modeling of failure probability and statistical design of SRAM array for yield enhancement in nanoscaled CMOS" IEEE Trans. on CAD, 24(12), 2005, pp. 1859-1880
[19] G. Reisfeld et al. "VARAN: Variability Analysis for Memory Cell Robustness" Proc. of SPIE, vol. 6925, March 2008.
[20] J. Wang et al. "Statistical modeling for the minimum standby supply voltage of a full SRAM array" Proc. ESSCIRC 2007, pp.400-403
[21] R. Aitken et al. "Worst-Case Design and Margin for Embedded SRAM" Proc. DATE, 2007
[22] P. Zuber, M. Miranda, et. al, Statistical SRAM analysis for yield enhancement. Proc. DATE, 2010, pp.57-62.