

Sparse Regression Driven Mixture Importance Sampling for Memory Design

Maria Malik, Rajiv V. Joshi, *Fellow, IEEE*, Rouwaida Kanj, *Senior Member, IEEE*, Shupeng Sun, Houman Homayoun, and Tong Li

Abstract—In this paper, we present a sparse regression (SpaRe) model-based yield analysis methodology and apply it to memory designs with state-of-the-art write-assist circuitry. At the core of its engine is a mixture importance sampling technique which consists of a uniform sampling stage and an importance sampling stage. The proposed methodology allows for fast and accurate statistical analysis of rare fail events. In our approach, a SpaRe model is built using the uniform sampling stage data points obtained via circuit simulation (CktSim). Along with the model, an optimal threshold value is determined for proper pass/fail predict capability. The model and the threshold value are then used to predict the response in the importance sampling stage. This alleviates the need for CktSims in the latter stage and introduces significant speedup compared to fully CktSim-based approaches. The SpaRe model-based yield analysis is tested on a 14-nm FinFET SRAM design, and the results corroborate well with that of full CktSim-based yield analysis. The methodology is used to compare multiple state-of-the-art SRAM designs including selective boost and write-assist designs. The operating V_{\min} ranges and trends corroborate well with hardware measurements.

Index Terms—Design for manufacturing, integrated circuit (IC) design, memory, rare events, sparse regression (SpaRe), SRAM, statistical analysis.

I. INTRODUCTION

WITH technology scaling, process variations pose a serious challenge to the design and analysis of integrated circuit (IC) design [1]–[4]. IC designs generally integrate various circuit components and each component needs to be robust to process variations. Memory designs suffer most leaving serious implications on the chip yield, especially for low power design thereby posing further challenges for the operation of portable devices. To achieve high yield, the failure rate of an SRAM bit-cell must be less than 0.0001% [5], [6]. With such strict requirements of less than one failing part per million, statistical yield analysis methodologies have been

developed to address the problem of rare event estimation with high confidence [6]–[8]. Kanj *et al.* [6] propose mixture importance sampling (MixIS). Unlike Monte Carlo, MixIS avoids simulating too many samples in the success region, instead, it is designed to cover more samples in the critical tail regions of the performance distribution. It involves two sampling stages: a uniform sampling stage and an importance sampling stage.

Singhee and Rutenbar [7] present a statistical blockade method that filters sample points in the tail regions of the performance distribution and hence reduces the number of needed circuit simulations (CktSims) to build a performance metric tail cumulative distribution function. Dong and Li [8] rely on Gibbs sampling for rare event estimation. All these techniques are fully CktSims based.

Sparse regression (SpaRe) techniques have been developed [9]–[11], [13], [14] to address modeling circuit designs in the presence of variability. In this paper, we explore the integration of orthogonal matching pursuit (OMP) [10] method along with MixIS for fast and accurate yield analysis in the presence of rare fail events. Hence, we propose a SpaRe model-based yield analysis methodology. An important contribution of this method is to bypass hundreds to thousands of CktSims typically required in the importance sampling stage of MixIS. Thus, a SpaRe model is built using the uniform sampling stage points. The resulting model is efficient and employs only a few critical feature vectors. Most importantly, it can accurately predict the failure points of the importance sampling stage with significant speedup compared to the pure CktSim-based yield analysis counterpart that relies fully on CktSims [6]. Accordingly, SpaRe reduces the required number of CktSims approximately by half compared to CktSim by completely eliminating the need for CktSims in the importance sampling (second) stage of MixIS.

Due to the nature of functional fails, the circuit response does not evaluate in the fail region. Hence, while the response is typically continuous in the pass region, it suffers from a discontinuity at the fail boundary; thus, it is typically represented by a single value in the fail region indicating that the cell fails to function properly. Henceforth, we will refer to this response data as discontinuous data due to the discontinuity at the fail boundary. For purposes of yield estimation, there is a need for proper pass/fail prediction. Hence, it is required to map the continuous SpaRe model-based response to the real response data. To enable this mapping, we determine, during the model building phase, and using the uniform stage data an optimal threshold value that minimizes the number of

Manuscript received February 26, 2017; revised July 15, 2017; accepted August 29, 2017. Date of publication September 28, 2017; date of current version December 27, 2017. (*Corresponding author: Maria Malik.*)

M. Malik and H. Homayoun are with George Mason University, Fairfax, VA 22030 USA (e-mail: mmalik9@gmu.edu; hhomayou@gmu.edu).

R. V. Joshi is with the IBM T. J. Watson Research Center, Yorktown Heights, NY 10598 USA (e-mail: rvjoshi@us.ibm.com).

R. Kanj is with the American University of Beirut, Beirut 1107 2020, Lebanon (e-mail: rk105@aub.edu.lb).

S. Sun is with Carnegie Mellon University, Pittsburgh, PA 15213 USA (e-mail: shupengs@ece.cmu.edu).

T. Li is with STG, IBM, Austin, TX 78758 USA (e-mail: tongli@us.ibm.com).

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TVLSI.2017.2753139

1063-8210 © 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.

See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

false predicts for the developed model. We rely on the model and its corresponding optimal threshold value to accurately predict passing or failing sample points for the importance sampling stage. This paper presents for the first time the application of such an accelerated fast statistical analysis tool that incorporates modeling cell functional fails such as write ability. It is implemented in the context of state-of-the-art selective boosting with write-assist circuitry. Selective boosting is applied to the memory and part of the logic virtual supply using “single supply.” The write-assist technique helps address the quantization of FinFETs that would otherwise pose problems due to improper cell beta and gamma ratios. The methodology, as well as the algorithms, uniquely pinpoint the advantages of the write-assist circuit technique.

This paper is organized as follows. Section II presents a review of MixIS and OMP. Section III presents the proposed methodology. Section IV presents memory designs under study. Section V presents the simulation analysis and results. Finally, conclusions are presented in Section VI.

II. BACKGROUND REVIEW

In this section, we present a review of MixIS methodology and OMP. This enables mapping the MixIS stages to the SpaRe model training and evaluation phases presented in the following section.

A. Mixture Importance Sampling Review

In order to guarantee high chip yield, it is essential for yield analysis tools to estimate low failure probabilities with good confidence. This requires a very large number of sample points using the traditional Monte Carlo methods. Importance sampling is a variance reduction method that focuses on generating more sample points in the critical fail regions [6]. Hence, instead of sampling, using the natural probability density function (PDF) $f(x)$, one would sample a distorted PDF $p(x)$ that biases the sampling to the important region, typically toward the tails of the distribution. In this case, the failure probability P_f can be derived as

$$P_f = \int_{-\infty}^{+\infty} \frac{I(x) \cdot f(x)}{p(x)} \cdot p(x) dx \quad (1)$$

where $I(x)$ is the indicator function

$$I(x) = \begin{cases} 1, & \text{fail } x \in R_F \\ 0, & \text{pass } x \notin R_F \end{cases} \quad (2)$$

and R_F denotes the failure region

The proper choice of $p(x)$ is critical. $p(x)$ is typically derived from the natural PDF $f(x)$ [6]. To address this problem, MixIS has been proposed in [6] to generate random variables using mixtures of distributions. The algorithm first performs uniform sampling and identifies the corresponding failure points. Next, the algorithm determines μ_s , the center of gravity of the failure points, and uses $p(x) = f(x - \mu_s)$ to generate the importance sample points according to [6] for the second stage of sampling. The true yield is then computed by unbiasing the estimate using weights representing the ratio of the natural to the distorted PDF according to (1).

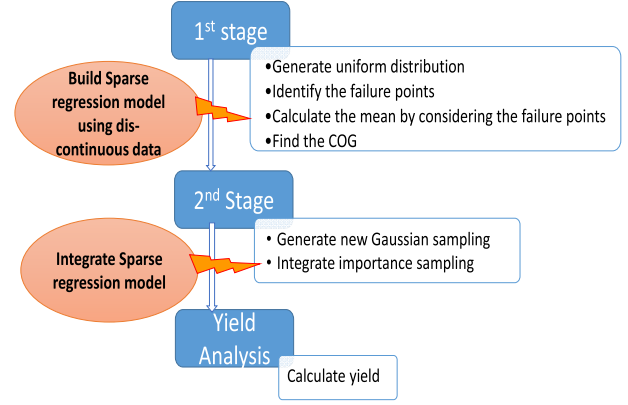


Fig. 1. Overview of SpaRe. The first stage and second stage represent the uniform and importance sampling stages of MixIS. SpaRe uses the first stage simulations to build a SpaRe model and predict the response of the second stage sample points.

B. OMP Overview

OMP methodology determines the critical unknown model coefficients via moment matching or inner product. For the case of a simple linear regression model of the form

$$y = \sum_i x_i * \beta_i + C \quad (3)$$

where y represents the response variable, and x_i represent the explanatory independent random variables, it can be shown that

$$\begin{aligned} E(y) &= C \\ E(y \cdot x_i) &= \beta_i. \end{aligned} \quad (4)$$

Hence, if the inner product ‘ $y \cdot x_i$ ’ (β_i) is far away from zero, then it is significant and x_i should be included in the model. The above property (4) can be extended to nonlinear functions of the form

$$y = \sum_i g(x_i) * \beta_i + C \quad (5)$$

by relying on orthogonal polynomial basis functions g_k , such as Hermite polynomials, [10], [15] satisfying the following relation:

$$\begin{aligned} E(g_i \cdot g_j) &= \begin{cases} 1 & \text{if } i = j \\ 0 & \text{o.w.} \end{cases} \\ E(y \cdot g_i(x)) &= \beta_i. \end{aligned} \quad (6)$$

III. PROPOSED METHODOLOGY

In this section, we present the proposed methodology overview and implementation details.

A. Methodology in a Nutshell

Fig. 1 presents an overview of our proposed methodology. It employs both CktSims and SpaRe models for the yield analysis engine, and involves two major steps.

- 1) Use the first stage uniform sample points simulations to build a SpaRe model and to calculate μ_s .
- 2) Use the SpaRe model to predict the failure sample points of the importance sampling stage, i.e., second stage, of MixIS.

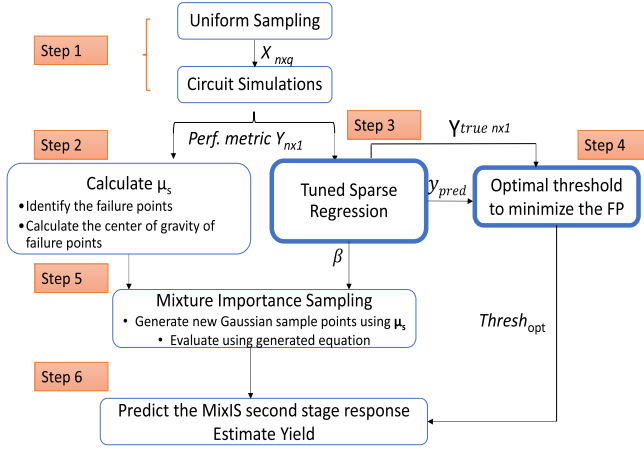


Fig. 2. SpaRe methodology steps flow diagram.

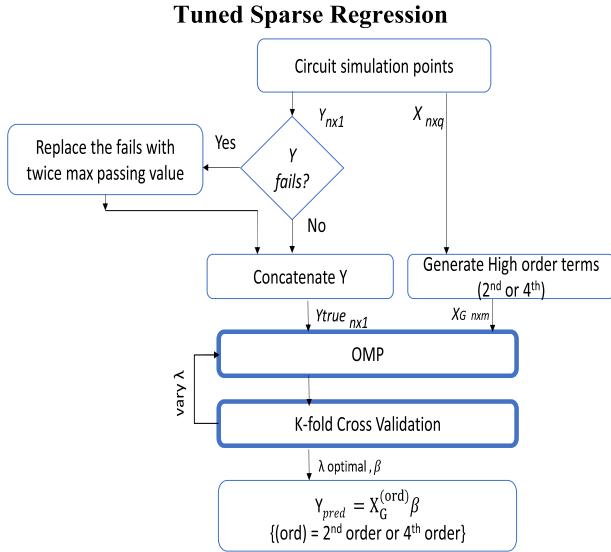


Fig. 3. Tuned SpaRe flow diagram. It represents step 3 of Fig. 2. Higher order polynomials include interaction terms.

B. Methodology Flow

Given q explanatory variables vector $X = \{x_1 \dots x_q\}$, and a response variable y . Our objective is to efficiently estimate the yield by employing SpaRe models. Fig. 2 presents the proposed methodology steps and flow diagram. The Methodology can be best described as follows.

Step 1: Generate and simulate n sample points uniformly over the q explanatory variables space.

Step 2: Find μ_s , and use it to shift the natural distribution of the explanatory variables to guarantee that the next stage importance sample points have good coverage of the region of fails.

Step 3: Perform tuned SpaRe for the uniform sample points using OMP and k -fold cross-validation technique as indicated in Fig. 3, and detailed in the following.

Given: X : $n \times q$ dimensional explanatory variables matrix

Y : $n \times 1$ dimensional response variable vector

Output: β : $m \times 1$ dimensional model coefficient vector

C : a model scalar entity

m represents the dimensionality of the derived higher order polynomial feature vector as will be explained in the following.

Step 3.a: As noted earlier, the circuit response does not evaluate in the failure region R_F . We typically return a constant value indicating failure instead. Hence, the data is continuous in the pass region and lumped at a single value indicating fail in the failure region. For our model building purposes, we define Y_{true} to be the vector derived from the original response vector Y as follows:

$$Y_{true} = \begin{cases} Y & Y \notin R_F \\ 2 * \max * (Y \notin R_F) & Y \in R_F \end{cases} \quad (7)$$

Hence, we modify the original response variable vector (Y) obtained from CktSim, and replace all the failure sample points with twice the maximum passing value; for example, in the case of write ability, we replace it by twice the maximum write delay. This provides separation between the pass and fail regions and allows room for error due to polynomial model fluctuations. Thereby, this enables safety margin for the threshold to separate between the pass and fail regions in the approximate continuous model.

Step 3.b: As discussed in Section II-B, we rely on the higher order polynomial functions for increased model accuracy and to enable the model to approximate well the complex nonlinear relationship between the explanatory and response variables. Thus, this step generates the high order $n \times m$ dimensional feature matrix

$$X_G = \{g_1(X), g_2(X), \dots, g_m(X)\} \quad (8)$$

where g_k are the Hermite polynomial basis functions that include interaction terms and higher order polynomial terms of the explanatory variables. For a simple 2-D case, the Hermite Polynomial is represented as [10]

$$\begin{aligned} g_1(x_i, x_j) &= 1 & g_2(x_i, x_j) &= x_i \\ g_3(x_i, x_j) &= \frac{1}{\sqrt{2}}(x_i^2 - 1) & g_4(x_i, x_j) &= x_i * x_j, \dots \end{aligned} \quad (9)$$

For higher dimensions, it can be derived according to [15]. For our purposes, our aim is to fit the model of the form

$$Y_{pred} = \sum_{i=1}^m X_{G,i}^{(ord)} \beta_i \quad (10)$$

where (ord) = second order or fourth order, and $X_G^{(ord)}$ is the corresponding extended feature matrix.

Step 3.c: Run the OMP [10] SpaRe with cross validation. OMP is used to solve the regularization problem of form (11), and accordingly identify the set of critical model features. Thereby producing a high-fidelity SpaRe model with a few important nonzero coefficients β derived according to (4). The input to OMP is Y_{true} and the feature matrix X_G

$$\min_{\beta} \|X_G \cdot \beta - Y_{true}\|_2^2 \text{ such that } \|\beta\|_0 < \lambda \quad (11)$$

where λ is the regularization parameter. OMP solves the otherwise NP-hard L0-norm regularization problem. Traditionally, the L0-norm problem is replaced by L1-norm problem, and the solution may involve costly optimizations. OMP finds the optimal number of features iteratively more efficiently [11], [12]. The optimal number of features is determined based on the



Fig. 4. Fivefold cross validation.

model corresponding to the lowest k -fold cross-validation error. Thus for a given model (set of features), and for a data set of size n , the data set is divided into k folds. A single fold is considered as the test set while remaining $k - 1$ folds are held out for training sets. We repeat this k times each time selecting a different fold for the test set to find the average cross-validation error. There are tradeoffs for the choice of k . Large k values are associated with a reduced bias to overestimating the model error; this comes at the expense of higher runtimes compared to small k values. On the other hand, small k values are typically associated with reduced model stability due to the increased size of the perturbation [16]. Hence the stability for a 20-fold cross validation is better than that of tenfold which is better than that of the hold-out method (train on 2/3 folds, and test on 1/3 [16]). For our purposes, our objective is to compare the average error among different models, and at the same time maintain a low runtime overhead due to the iterative nature of OMP feature selection process. Without loss of generality, we rely on fivefold cross validation. This has a relatively low runtime overhead for purposes of the L0-norm regularization, and helps us compare the different models' errors and hence identify the optimal number of features corresponding to minimum cross-validation error efficiently. We thus divide the input data set into five mutually exclusive folds, and run the cross-validation process 5 times as illustrated in Fig. 4.

OMP can be best summarized by the following steps.

- 1) Initialize the residual to be $R = Y_{\text{true}}$; and the set of selected features to be $\Omega = \{\}$. Set the loop index $l = 1$.
- 2) Compute correlation between R and the features $X_{G,i}$

$$c_i = \langle R, X_{G,i} \rangle.$$

- 3) Select from $\{X_{G,i}\} - \Omega$ the feature k with the highest correlation to R . update $\Omega = \Omega + \{X_{G,k}\}$.
- 4) Build the linear regression model for Y_{true} using only the selected features in Ω

$$\min_a \left\| \sum_{i \in \Omega} X_{G,i} \cdot a_i - Y_{\text{true}} \right\|_2^2.$$

- 5) Calculate the corresponding average k -fold cross-validation error e_l .
- 6) Update R based on the new model, such that

$$R = Y_{\text{true}} - \sum_{i \in \Omega} X_{G,i} \cdot a_i + C.$$

7) If $\text{card}(\Omega)$ did not exceed λ , $l = l + 1$ and go to step 2. Otherwise go to step 7.

8) Use the model corresponding to the iteration with minimum cross-validation error e_l . Set the critical β values to match α values of the best models. Set the remaining β coefficients to 0 for $X_{G,i}$ that are not selected in that model.

At each new iteration, the methodology adds to the set of critical features the feature that portrays the highest correlation with the residual obtained from the previous iteration model. A model is built with new feature set, a new residual is computed, and the corresponding cross-validation error is computed. Typically, we stop the search once the number of features exceeds a certain maximum limit λ_{max} predefined by the user (for example 50 features at max). Finally, the model with minimum cross-validation error is adopted.

Step 4: The last step of model building involves identifying a threshold value that properly maps the true pass/fail points of the uniform data sample points to those predicted by the model (Fig. 5). Any predicted sample points (Y_{pred}) greater than threshold value are considered as a fail, and our ideal objective would be to satisfy the following relation:

$$Y_{\text{pred}} > \text{thresh}_{\text{opt}} \Rightarrow Y_{\text{true}} \in R_F.$$

In practice, we find for the optimal threshold $\text{thresh}_{\text{opt}}$ satisfying the following relations:

$$\min_t \text{FalsePredict} \quad (12)$$

where $t \in [\min(Y_{\text{pred}}), \max(Y_{\text{pred}})]$ is a threshold value selected over the range of Y_{pred} values

$$\text{FalsePredict} = \text{FalseNegative} + \text{FalsePositive}$$

$$\text{FalseNegative} = \sum (Y_{\text{pred}} < t) * I(Y_{\text{true}})$$

$$\text{FalsePositive} = \sum (Y_{\text{pred}} > t) * (1 - I(Y_{\text{true}}))$$

and $I(Y_{\text{true}})$ is an indicator function that evaluates to one when Y_{true} belongs to the failure region similar to (2). To find the optimum t , we sweep it over all Y_{pred} values for the specific set of training sample points. Fig. 6 presents the pseudocode for searching for the optimal threshold given the training set of n uniform stage sample points. Note that rather than repeating the summations in (11) for every new threshold value, we rely on an incremental update for FalseNegative and FalsePositive values. In fact, the number of false negatives for a given threshold value can be derived incrementally from the previous value obtained prior to increasing the threshold value. The number of false positives holds an opposite relation, and hence, we derive it backwards. Maximum threshold implies no predicted fails and hence no false positives.

In the results section, we discuss using an upper bound on the false negatives to avoid being overly optimistic. The optimal threshold is passed along with model coefficients for predicting the response of the importance sample points of the second MixIS stage.

Step 5: Generate importance sample points using the shifted Gaussian (natural) distribution [6], and estimate the performance metric and predict failures using the model coefficients,

Optimal threshold to minimize the FP

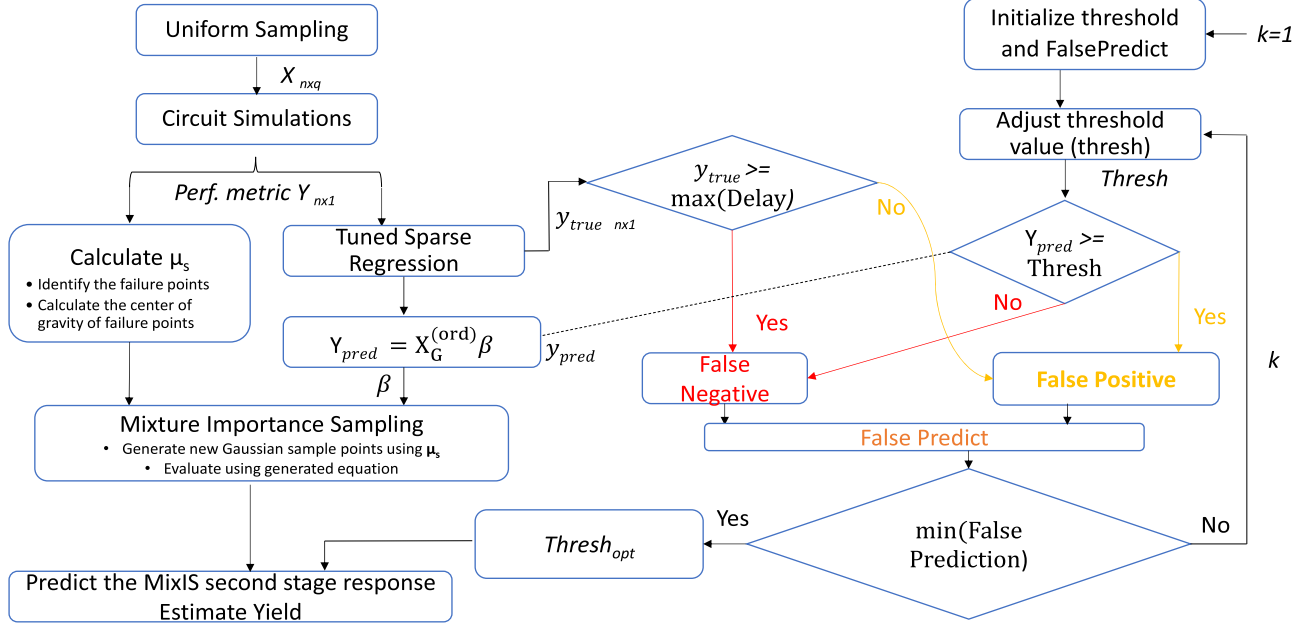


Fig. 5. Calculating the optimal threshold to minimize false prediction. This represents step 4 of Fig. 2.

Finding Threshold Opt

```

Ypred_sort ← sort(Ypred), Define I_dx | {Ypred_sort,i = Ypred_I_dx(i)}
Ytrue_sort,i = Ytrue_I_dx(i)
/***** Initialize *****/
minFalsePredict=1e6;
t(1) = Ypred_sort,1;
/***** False Negative Loop *****/
falseNegative(1) = I(Ytrue_sort,1)
For i=2:n
    t(i) = Ypred_sort,i;
    falseNegative(i) = falseNegative(i-1) + I(Ytrue_sort,i)
End
/***** False Positive Loop *****/
falsePositive(n) = 1 - I(Ytrue_sort,n)
For i=n-1:1
    falsePositive(i) = falsePositive(i+1) + ...
                    ... (1 - I(Ytrue_sort,i))
    falsePredict(i) = falseNegative(i) + falsePositive(i)
    If (falsePredict(i) < minFalsePredict)
        minFalsePredict = falsePredict(i)
        i_opt = i
        thresh_opt = t(i)
    End
End
End

```

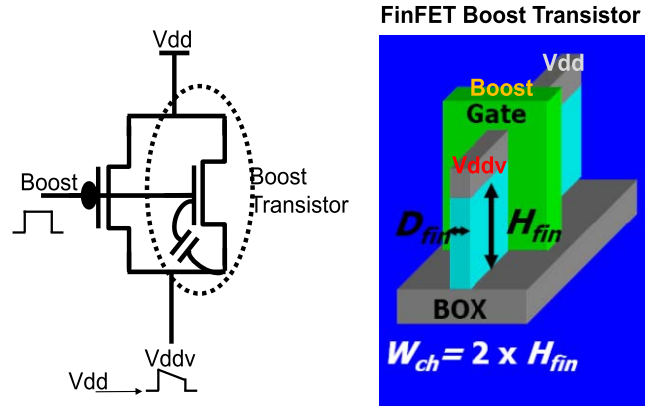
Fig. 6. Pseudo code for finding the optimal threshold for n sample points.

β as generated in step 3 and optimal threshold value discussed in step 4.

Step 6: Estimate the probability of fail and compute the corresponding yield using equations similar to (1).

IV. APPLICATION TO STATE-OF-THE-ART WRITE-ASSIST CIRCUITRY

Nonplanar technology (FinFET) brings forth a new challenge in terms of the quantization of FinFETs which poses a problem for proper beta and gamma ratios used in the cell. A minimum sized SRAM cell with 1 fin each for all the devices drives a beta and gamma ratio of 1. This disturbs

Fig. 7. Capacitive coupling between gate and source boosts the source voltage (V_{ddv}) above V_{dd} when boost switches “high” [17].

the “write ability” at low “ V_{dd} .” To overcome such problems new circuit techniques are essential. In this paper, we evaluate unique selective transistor-based boosting circuit techniques along with write-assist techniques for the purpose of improving low voltage operation.

Fig. 7 presents a typical boost circuit. It involves an nFET in parallel with a pFET device. The pFET source and nFET drain are connected to V_{dd} . When their common gate is switched from low to high, it couples the nFET’s source (pFET’s drain) to a value above V_{dd} . The devices are sized to give a boost around 0.12 V at low voltage operation. The selective boost technique allows the design to operate at lower V_{dd} values by selectively boosting memory specific paths and excluding surrounding logic from the boost. The selective boosting can be further paired with other write-assist techniques.

Thus, selective boost when applied to selective path-wordline, write drivers and cell, therefore, helps improve the yield and pushes the low voltage operation range. The write-assist techniques further help the SRAM cells improve

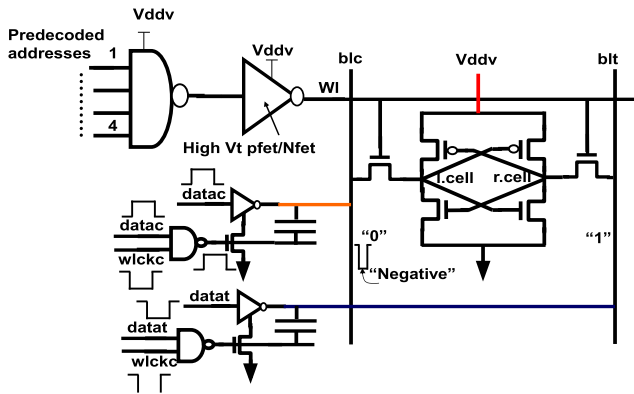


Fig. 8. Negative boost write assist—applied to bitlines with virtual supply (V_{ddv}) [17].

TABLE I
DESIGNS UNDERSTUDY

Design	Write Assist	Referred as
Selective Boost Only	None	SB_NoAssist
Selective Boost with Write Assist	Voltage Collapse	SB_Collapse
	Negative Bitline Boost	SB_NegBoost

the write-ability yield. Two such techniques are: the voltage collapse and negative bitline boost methods [16]. The earlier operates by lowering the cell voltage during write. The latter [16] relies on negative boosting of the bitlines (Fig. 8). The negative boost is created through capacitor coupling between the gate nodes of *datat* (data true) and *datac* write enable transistors and the bitlines. During the write operation, the negative boost brings the bitline voltage lower than zero and generates an increased voltage swing between the true and complement bitline during write. Thus, the increase at the gate voltage makes the transistor strong so it can flip the cell bit easily and enhance the write ability.

For near threshold voltages, the devices are typically weak and selective boosting is needed for both read and write operations. In fact, boosting further amplifies the write assist improvements [17]. For higher voltages, boost is turned on only for read operation. Finally, programmable booster designs [18] can further enhance read access times by properly tuning boost signal pulsewidth and phase.

Table I lists three design options that were implemented in [17] for a 14-nm FinFET SOI technology 72-Kb SRAM array arranged in columns of 16 cells/bitline. The presented techniques are as follows:

- 1) selective boosting only;
- 2) selective boosting paired with voltage collapse;
- 3) selective boosting paired with negative bitline boost.

Hardware measurements indicate that SB_NoAssist design requires more than 0.45 V to operate. SB_Collapse allows the cells to work with the low voltage as of 0.35 V. SB_NegBoost further stretches the operating voltage range all the way to 0.30 V for write ability. These V_{min} operating ranges and trends were designed and validated using statistical design methodologies [6] and results were found to corroborate well with fabricated hardware presented in Fig. 9. Fig. 10 illustrates the yield improvement due to SB_Collapse and SB_NegBoost

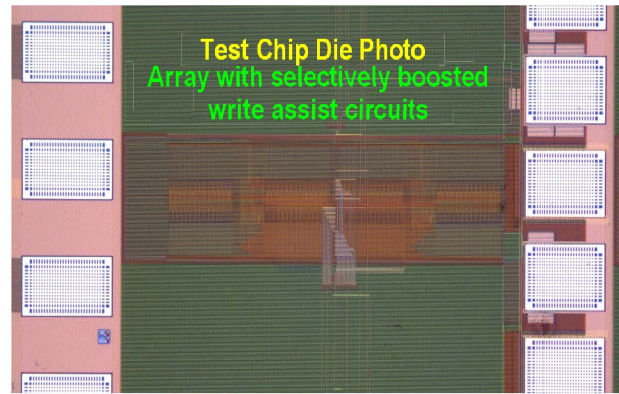


Fig. 9. Die photograph in 14-nm FinFET SOI technology.

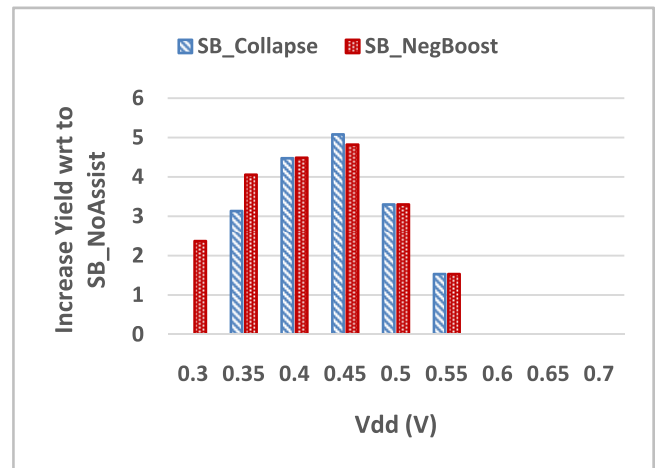


Fig. 10. Increase in yield with SB_Collapse and SB_NegBoost compared to the SB_NoAssist (arbitrary units).

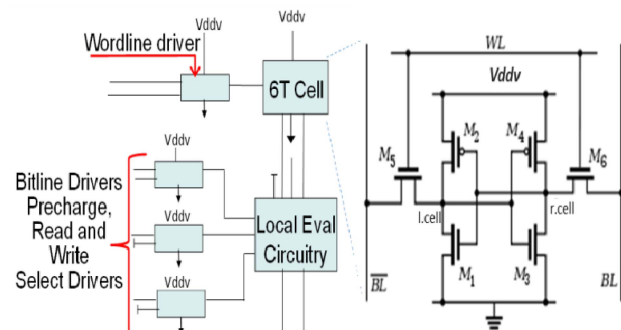


Fig. 11. SRAM cell cross section. 6 T portion of 8 T cell illustrated.

compared to the SB_NoAssist. It is clear that write assist SB_NegBoost provides significant yield improvement compared to SB_Collapse as it can operate at the lower supply voltage.

V. ANALYSIS AND RESULTS

For purposes of our analysis, we demonstrate the efficacy of the proposed SpaRe methodology in the evaluation of the negative bitline boost write-assist (SB_NegBoost) technique in terms of model prediction capability and yield estimation.

A. Experimental Setup

We apply the methodology to an industrial 14-nm FinFET SRAM design. For accurate analysis, simulations involve the cell along with the peripheral logic as illustrated in Fig. 11.

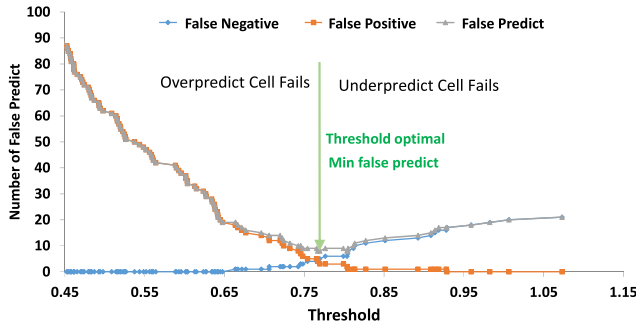


Fig. 12. Optimal threshold selection. Small threshold values overpredict cell fails. Large threshold values underpredict cell fails.

Variability is injected in the memory cell devices as well as the local evaluation circuitry. Variability effects such as metal gate granularity, line edge roughness, fin height variation, and random dopant fluctuations are lumped into one source σ_{vt} that is injected into the simulations. Write ability, which is the ability of the cell to be written, is selected as the primary metric for investigation.

B. Model Building: Uniform Sampling Stage Data

We apply the proposed methodology to analyze the designs under study as the supply voltage is varied over the range [0.40–0.43 V] for the SB_NoAssist design and [0.35–0.40 V] for the SB_NegBoost design. For each design point, we generated 1000 samples using uniform distribution for nine features/explanatory variables. Due to the nonlinearity of memory designs, we build a higher order polynomial model (second order or fourth order) to improve the accuracy. To handle sparsity, OMP expresses the model as the function of few important nonzero coefficients β . For instance, with the fourth order polynomial model, OMP focuses only on a maximum of 50 nonzero coefficients β as compared to a full blown model with 714 terms. This is based on the fact that for k variables, the number of expansion coefficients for upto n th degree polynomial is derived according to $\binom{k+n}{n} - 1$ [19]. Finally, we employ a fivefold cross-validation approach for optimal λ .

C. Optimal Threshold Selection Analysis

We determine the threshold value used to predict the pass and fail criteria for the importance sampling stage of MixIS as the value that minimizes the uniform sampling stage false predictions (see Figs. 2 and 5). Fig. 12 presents an example optimal threshold value Thresh_{opt} . We observe that for small threshold values the model is pessimistic and tends to overpredict cell fails due to a high number of false positive; small threshold implies a lot of Y_{pred} values fail. For large threshold values, the model is optimistic, tends to underpredict the number of fails due to an increase in false negatives. We select the optimal threshold value that provides the minimum false predict. For all our experiments, the model corroborated well and displayed a proper monotonic trend with the true data. At the optimal threshold value, we had on average 1.5% false predictions for the uniform stage sample points training data set. To guard against overly optimistic scenarios should

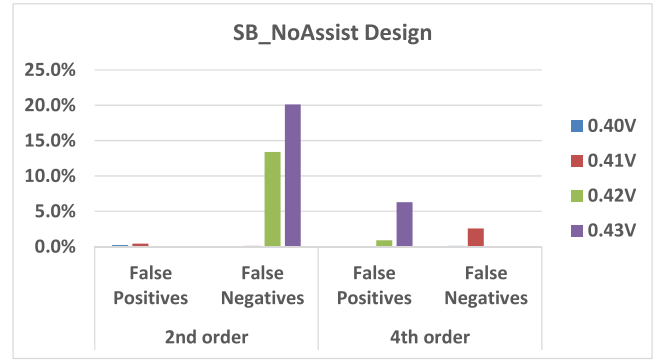


Fig. 13. SpaRe prediction summary for MixIS second stage for SB_NoAssist design. Golden simulations are based on CktSim.

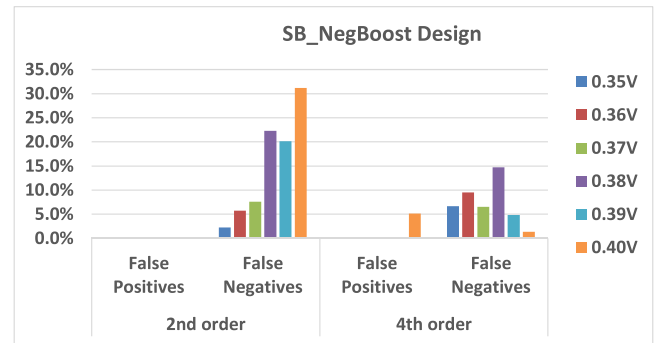


Fig. 14. SpaRe prediction summary for MixIS second stage for SB_NegBoost design. Golden simulations are based on CktSim.

they arise, we put a limit in our code on the maximum tolerable false negatives to be 10% of the number of fails detected in the uniform sampling stage. Hence, we adjust the threshold accordingly if needed.

D. Model Prediction: Importance Sampling Stage

In this section, we focus on evaluating the proposed model prediction capabilities in comparison to pure CktSims for the sample points of the importance sampling stage of MixIS. We present the percentage of false positives and false negatives (uncaptured fails) for both designs in Figs. 13 and 14, respectively. The results are for second order and fourth order polynomial models. In the following section, we evaluate the corresponding yield convergence of the proposed methodology.

Overall, we find that the fourth order model has better prediction capability for both designs. We report the results in terms of the percent false positives and percent false negatives compared to the true number of fail points. For the second order polynomial, SB_NoAssist design, we observe that false positives percentage is as small as 0.4% for all the voltages. The maximum error is 20% for the false negative at 0.43-V supply voltage. On the other hand, the maximum error for the fourth order polynomial is 6% recorded for the false positives at 0.43 V.

For the SB_NegBoost design, again the fourth order polynomial demonstrates lower errors in comparison to the second order. The maximum recorded error is 15% for the fourth order model false negatives recorded at 0.38 V.

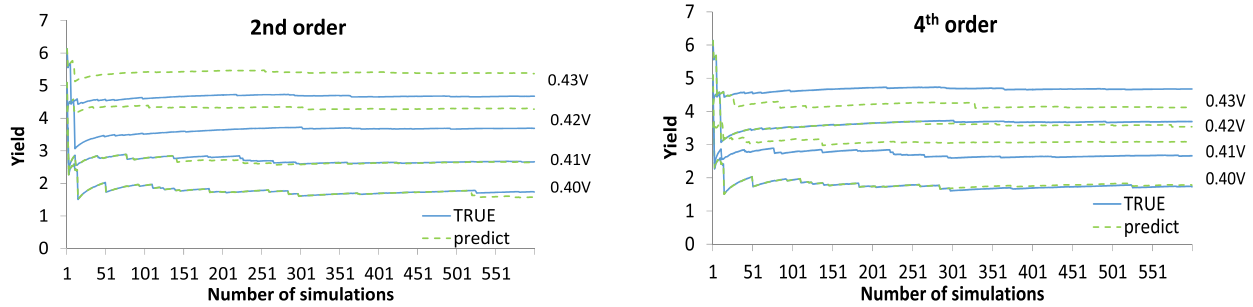


Fig. 15. Yield estimation: SB_NoAssist design-convergence. TRUE = CktSim. Predict = SpaRe.

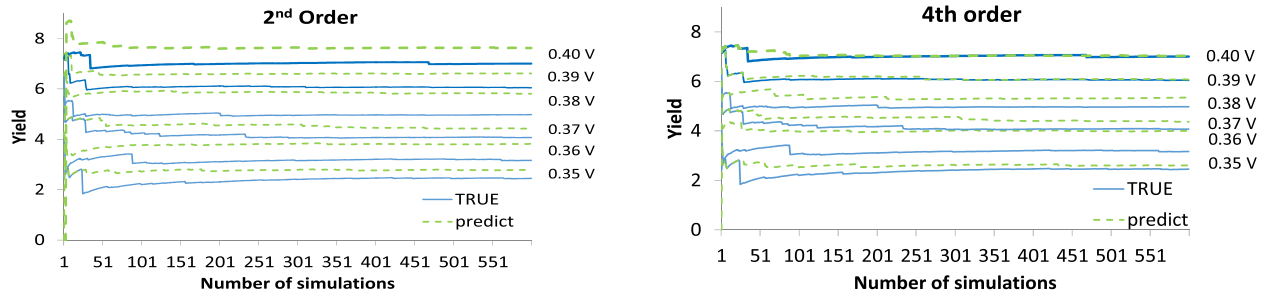


Fig. 16. Yield estimation: SB_NegBoost design-convergence. TRUE = CktSim. Predict = SpaRe.

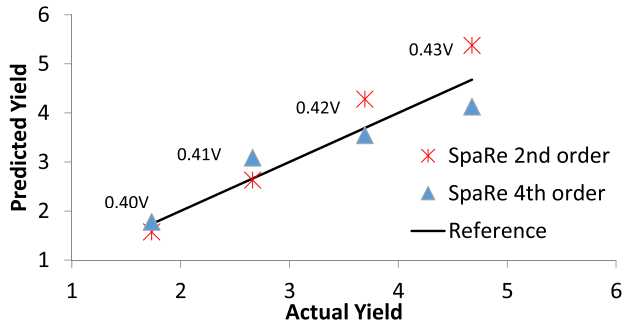


Fig. 17. SB_NoAssist design yield estimation. Arbitrary units. Actual yield = CktSim. Predicted yield = SRI.

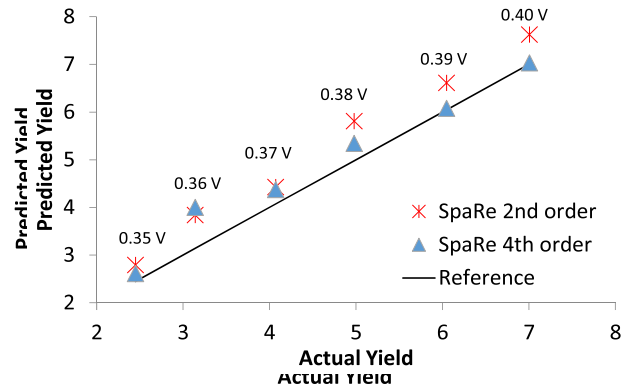


Fig. 18. SB_NegBoost design yield estimation. Arbitrary units. Actual yield = CktSim. Predicted yield = SpaRe.

E. Yield Estimation and Convergence Analysis

We apply the SpaRe methodology to estimate the yield of the SB_NoAssist and SB_NegBoost memory design. Hence, we rely on the model built using the uniform sampling stage sample points of MixIS (first stage), to predict the response for the sample points generated in the importance sampling stage of MixIS (second stage). This completely eliminates the need for CktSim in the importance sampling stage. We analyze the convergence of the yield estimate for both the proposed SpaRe methodology against CktSim, the pure CktSim-based approach for MixIS. For our proposed technique, the yield estimation is based on the second order and fourth order polynomial models presented in the previous section. The respective yield convergence results are presented in Figs. 15 and 16. We observe that for both SB_NoAssist design and SB_NegBoost design, second order polynomial does not provide proper yield convergence especially at the high supply voltages. This is expected due to the higher errors observed for the second order model prediction presented in Section V-D. However, the fourth order model yield estimation results demonstrate high corroboration between

the proposed SpaRe methodology and the traditional CktSim methodology for both the SB_NoAssist and SB_NegBoost designs. Note that for the fourth order model-based analysis, the maximum yield error was 15.9%, and the average error was found to be 6.2%. This is emphasized in Figs. 17 and 18 where the final converging values of the second order and fourth order model-based yield estimates are compared. Specifically, we observe the following for the fourth order model.

- 1) For SB_NoAssist design yield prediction properly matches the response at 0.4 and 0.42 with a slight mismatch at 0.41 and 0.43 V.
- 2) For the SB_NegBoost design, the predicted yield results match properly with slight mismatch at 0.36 V.
- 3) For both designs, the predicted yield trend is preserved and the yield is monotonically increasing with the voltage increase.
- 4) Most importantly, Fig. 19 illustrates the ability of the fourth order model-based to accurately predict the corresponding low fail probabilities.

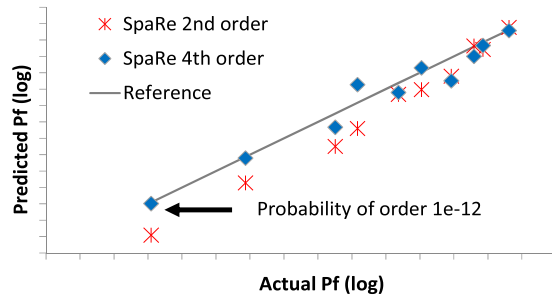


Fig. 19. Probability of fail estimation corresponding to Figs. 17 and 18.

TABLE II
RUNTIME COST

	CktSim	SpaRe
Uniform Sampling Stage	Circuit simulations 5 hours	Circuit Simulations 5 hours
Model fitting	N/A	110 seconds
Importance Sampling stage	7.5 hours	< 1 min

As far as the runtime is concerned, the proposed SpaRe methodology reduces the required CktSims by half and completely eliminates the need of CktSim for the second stage of mixture important sampling. The methodology can be effectively applied to other chip design frameworks in general. It enables significant speedup compared to the traditional CktSim-based techniques which can be very costly for large cross sections. Furthermore, there is no added cost for training sample points since these are obtained in the uniform sampling stage. We can think of OMP as least squares regression repeated iteratively. Because OMP solves least squares based on a small number of features, $\lambda_{\max} = 50$, this implies that the complexity of matrix computation and inversion required for least squares regression is reduced significantly due to the reduced number of features. From a practical perspective, we recorded around 110 s of runtime for the model building (fitting) using MATLAB. The model evaluation is negligible. This is compared to more than 7.5 h runtime that was required for CktSims of the importance sampling stage as illustrated in Table II. This accounts for 150 \times reduction in runtime. Both MATLAB and SPICE simulations were performed on an IBM Power7 core processor machine running at 4 GHz.

VI. CONCLUSION

We present a first application of a SpaRe model-based yield analysis methodology for rare fail estimation of memory designs. In the proposed methodology, we integrate the SpaRe model using an optimal threshold value to predict the failures of the importance sampling stage of MixIS for yield analysis calculations. The methodology is shown to efficiently model cell functionality in the context of high-dimensional space with application to state-of-the-art selective boosting with write-assist circuitry. The methodology bypasses hundreds to thousands of CktSims by completely eliminating the need for CktSims of the second phase of MixIS. This accelerates the statistical analysis of memory designs without compromising accuracy.

REFERENCES

- [1] V. De and S. Borkar, "Technology and design challenges for low power and high performance [microprocessors]," in *Proc. Int. Symp. Low Power Electron. Design*, Aug. 1999, pp. 163–168.
- [2] S. R. Nassif, "Design for variability in DSM technologies [deep sub-micron technologies]," in *Proc. IEEE 1st Int. Symp. Quality Electron. Design (ISQED)*, Mar. 2000, pp. 451–454.
- [3] R. V. Joshi *et al.*, "A low power and high performance SOI SRAM circuit design with improved cell stability," in *Proc. IEEE Int. SOI Conf.*, Oct. 2006, pp. 4–7.
- [4] C. Visweswariah, "Plenary speech 2P2: Statistical techniques to achieve robustness and quality," in *Proc. 9th Int. Symp. Quality Electron. Design (ISQED)*, Mar. 2008, p. 586.
- [5] S. Sun, X. Li, H. Liu, K. Luo, and B. Gu, "Fast statistical analysis of rare circuit failure events via scaled-sigma sampling for high-dimensional variation space," in *Proc. Int. Conf. Comput.-Aided Design*, Nov. 2013, pp. 478–485.
- [6] 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," in *Proc. 43rd ACM/IEEE Design Autom. Conf.*, Jul. 2006, pp. 69–72.
- [7] A. Singhee and R. A. Rutenbar, "Statistical blockade: A novel method for very fast Monte Carlo simulation of rare circuit events, and its application," in *Proc. Design, Autom. Test Eur. Conf. Exhib. (DATE)*, 2007, pp. 235–251.
- [8] C. Dong and X. Li, "Efficient SRAM failure rate prediction via Gibbs sampling," in *Proc. 48th ACM/EDAC/IEEE Design Autom. Conf. (DAC)*, Jun. 2011, pp. 200–205.
- [9] X. Li, J. Le, and L. T. Pileggi, *Statistical Performance Modeling and Optimization* (Foundations and Trends in Electronic Design Automation), vol. 1. Boston, MA, USA: Now Publishers Inc., 2006, pp. 331–480. [Online]. Available: <http://dx.doi.org/10.1561/1000000008>
- [10] X. Li and H. Liu, "Statistical regression for efficient high-dimensional modeling of analog and mixed-signal performance variations," in *Proc. DAC*, Jun. 2008, pp. 38–43.
- [11] X. Li, "Finding deterministic solution from underdetermined equation: Large-scale performance modeling by least angle regression," in *Proc. DAC*, Jul. 2009, pp. 364–369.
- [12] M. B. Alawieh, F. Wang, R. Kanj, X. Li, and R. Joshi, "Efficient analog circuit optimization using sparse regression and error margining," in *Proc. 17th Int. Symp. Quality Electron. Design (ISQED)*, Mar. 2016, pp. 410–415.
- [13] Y. Wang, M. Orshansky, and C. Caramanis, "Enabling efficient analog synthesis by coupling sparse regression and polynomial optimization," in *Proc. DAC*, Jun. 2014, pp. 1–6.
- [14] Y. Zhang, S. Sankaranarayanan, and F. Somenzi, "Sparse statistical model inference for analog circuits under process variations," in *Proc. ASP-DAC*, Jan. 2014, pp. 449–454.
- [15] G. Sansone, *Orthogonal Functions*. New York, NY, USA: Dover, 2004.
- [16] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proc. IJCAI*, 1995, vol. 14, no. 2, pp. 1137–1145.
- [17] R. V. Joshi, M. Ziegler, H. Wetter, C. Wandel, and H. Ainspan, "14 nm FinFET based supply voltage boosting techniques for extreme low V_{\min} operation," in *Proc. Symp. VLSI Circuits*, Jun. 2015, pp. C268–C269.
- [18] R. V. Joshi and M. M. Ziegler, "Programmable supply boosting techniques for near threshold and wide operating voltage SRAM," in *Proc. IEEE CICC*, Apr./May 2017, pp. 1–4.
- [19] S. Rahman, "Wiener–Hermite polynomial expansion for multivariate Gaussian probability measures," *J. Math. Anal. Appl.*, vol. 454, no. 1, pp. 303–334, 2017.

Maria Malik received the B.E. degree in computer engineering from the Center of Advanced Studies in Engineering, Islamabad, Pakistan, and the M.S. degree in computer engineering from the George Washington University, Washington, DC, USA. She is currently pursuing the Ph.D. degree with the Electrical and Computer Engineering Department, George Mason University, Fairfax, VA, USA.

Her current research interests include computer architecture with the focus of performance characterization and energy optimization of big data applications on the high-performance servers and low-power embedded servers, accelerating machine learning kernels, parallel programming languages, and parallel computing.

Rajiv V. Joshi (M'87–F'01) received the B.Tech. degree from IIT Bombay, Mumbai, India, the M.S. degree from MIT, Cambridge, MA, USA, and the Dr.Eng. Sc. degree from Columbia University, New York, NY, USA.

He is currently a Research Staff Member with the T. J. Watson Research Center, IBM, Yorktown Heights, NY, USA. He has authored or co-authored over 185 papers. He holds 58 invention plateaus, 225 U.S. patents, and over 350 including international patents.

Dr. Joshi is a member of IBM Academy of technology. He received the Best Editor Award from the IEEE TVLSI journal, and the 2013 IEEE CAS Industrial Pioneer award and the 2013 Mehboob Khan Award from Semiconductor Research Corporation. He was a recipient of the 2015 BMM Award. He is inducted into New Jersey Inventor Hall of Fame in 2014 along with pioneer Nikola Tesla. He served as a Distinguished Lecturer of the IEEE CAS and EDS society. He is an ISQED and World Technology Network Fellow and distinguished alumnus of IIT Bombay. He is on the Board of Governors for the IEEE CAS. He serves as an Associate Editor of the IEEE TVLSI. He served on committees of International Symposium Low Power Electronic Design, the IEEE VLSI design, the IEEE CICC, the IEEE International SOI Conference, ISQED, and Advanced Metallization Program committees.

Rouwaida Kanj (SM'12) received the M.S. and Ph.D. degrees in electrical engineering from the University of Illinois at Urbana–Champaign, IL, USA, in 2000 and 2004, respectively.

From 2004 to 2012, she was a Research Staff Member at IBM Austin Research Laboratory, Austin, TX, USA. She was involved in modeling SOI effects, noise characterization of CMOS circuits, library characterization of novel circuit technologies, and is currently involved in statistical design methodologies. She has authored of more than 55 technical papers, 20 issued patents, and several pending patents. She is currently an Assistant Professor with the American University of Beirut, Beirut, Lebanon.

Dr. Kanj was a recipient of the three IBM Ph.D. Fellowships, the IEEE/ACM WILLIAM J. MCCALLA ICCAD Best Paper Award in 2009, and the ISQED Best Paper Award in 2006 and 2014. She received an Outstanding Technical Achievement Award and six Invention Plateau awards from IBM.

Shupeng Sun received the B.E. degree from the Department of Automation, Tsinghua University, Beijing, China, in 2010, and the Ph.D. degree from the Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA, USA, in 2015.

He is currently a Software Engineer with Google Corporation, Mountain View, CA, USA.

Dr. Sun was a recipient of the IEEE Donald O. Pederson Best Paper Award in 2016, and the ACM SIGDA DAC Ph.D. forum Best Poster Research Award in 2014. He won the Gold medal in the ACM Student Research Competition at ICCAD in 2014, and second place in the ACM Student Research Competition grand finals in 2015.

Houman Homayoun received the B.S. degree in electrical engineering from the Sharif University of Technology, Tehran, Iran, in 2003, the M.S. degree in computer engineering from the University of Victoria, Victoria, BC, Canada, in 2005, and the Ph.D. degree from the Department of Computer Science, University of California, Irvine, CA, USA, in 2010.

He was with University of California, San Diego, CA, USA, where he was a National Science Foundation Computing Innovation Fellow awarded by the CRA and CCC. He is currently an Assistant Professor with the ECE Department, George Mason University (GMU), Fairfax, VA, USA. He also holds a joint appointment with the Computer Science as well as Information Science and Technology Departments. He is the Director of Green Computing and Heterogeneous Architectures Laboratory, GMU. He is currently leading a number of research projects, including the design of next generation heterogeneous multicore accelerator for big data processing, nonvolatile spin transfer torque logic, heterogeneous accelerator platforms for wearable biomedical computing, and logical vanishable design to enhance hardware securities which are all funded by National Science Foundation, General Motors Company, and Defense Advanced Research Projects Agency.

Tong Li received the bachelor's and master's degrees from Electrical Engineering Department, Xian Jiaotong University, Xi'an, China, in 1989 and 1992, respectively, and the Ph.D. degree from Electrical Engineering Department, New Jersey Institute of Technology, Newark, NJ, USA, in 1999.

Since 2004, he has been with IBM EDA, Austin, TX, USA, focusing on circuit simulation, fast transistor model generation, and yield analysis of circuit design.