IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

IEEE RELIABILITY SOCIETY SECTION

Received January 23, 2021, accepted February 14, 2021, date of publication March 19, 2021, date of current version March 31, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3067510

A Spline-High Dimensional Model Representation for SRAM Yield Estimation in High Sigma and High Dimensional Scenarios

LIANG PANG, SHAN SHEN, AND MENGYUN YAO
National ASIC System Engineering Research Center, Southeast University, Nanjing 210096, China
Corresponding author: Liang Pang (almostday0@gmail.com)

This work was supported in part by the National Key Research and Development Program of China under Grant 2019YFB6506000915.

ABSTRACT Traditional Static Random-Access Memory (SRAM) yield estimation through Monte Carlo analysis is an extremely time-consuming process since it runs millions of expensive transistor-level simulations to get the yield results with the specified precision, especially for the large-scale circuits. In this paper, we develop an efficient yield analysis framework by integrating our novel performance metamodel into a state-of-art importance sampling method. The performance meta-model, named Spline-High Dimensional Model Representation (SP-HDMR), is used to substitute the expensive transistor-level simulations in yield estimation. The proposed SP-HDMR model provides a high computationally efficient formula expansion. It uses spline functions as the kernels to describe the various relations between the process parameters and SRAM read access delay. And an adaptive sampling method with sparsity analysis is developed to support SP-HDMR modeling. The experiments on the 40nm SRAM circuits validate the accuracy and the efficiency of the proposed yield analysis framework based on our SP-HDMR model with 1.3X∼5X speedup over the other state-of-art methods within 9% relative error.

INDEX TERMS Yield analysis, SRAM, importance sampling, SP-HDMR model.

I. INTRODUCTION

As semiconductor technology continues to advance, SRAM cells designed with minimum sizes are more susceptible to process fluctuations [1]. As a result, yield degeneration has become a bottleneck for the robust SRAM design [1]–[4]. To guarantee a robust design, traditional corner-based analysis methods will lead to a too pessimistic result in the worst-case corner and have to be verified on thousands of corners if more process parameters are considered. Thus, statistical methods are required to make yield estimations reasonably realistic. An SRAM chip typically consists of millions of SRAM cells. To make sure the acceptable chip yield, the failure rate of each cell should be extremely low. For a 1Kbits SRAM chip with 99% yield, the failure rate of an SRAM cell should be lower than 10^{-5}.

To estimate the failure rate of SRAM accurately, many statistical methods have been proposed. Among them, Standard Monte Carlo (MC) analysis is the traditional method and is remained as the gold standard. It samples the whole variation space directly and simulates each sample to get the corresponding performance at the transistor-level. However, it is extremely time-consuming to estimate SRAM failure rates due to the huge number of simulations, e.g. it needs over 10^7 simulations to get a 4-sigma yield result. Besides, in an SRAM array, the dynamic functional failure, such as read delay failures defined as read operations exceeding a specified time in our work, depends on not only the state of the weakest cell but also the state of other cells in the same column [5]. Hence, we must consider the process variation of all these cells to estimate the SRAM yield. As a result, it brings a high dimensional variation space for SRAM yield estimation and extremely expensive computational overhead per simulation.

To accelerate the traditional MC method, many statistical approaches have been proposed, which can be grouped into two categories:

Importance Sampling (IS): The basic idea of IS is to find a distorted sampling distribution to sample near the failure region. The efficiency and accuracy of IS-based methods heavily depend on the distorted sampling distribution. Most of the approaches [6]–[17] have been proposed to construct

The associate editor coordinating the review of this manuscript and approving it for publication was Jian Guo.

such distribution by shifting the mean-vector of the original distribution to the boundary or center of the failure region. Recently, Shi and Liu *et al.* [6] proposed the Adaptive Importance Sampling (AIS) which improves the tolerance of poor initialization by searching the failure boundary dynamically in resampling iterations through a developed unbiased estimator. However, the methods above are infeasible in high dimensional scenarios because the likelihood ratios between the original sampling distribution and distort sampling distribution have huge numerical instability [18], [24]. Wu and Gong *et al.* [9] proposed a kind of High Dimensional Importance Sampling (HDIS) to address this issue by constructing a new subset to calculate the failure rate indirectly but it consumes a large number of transistor-level simulations to converge. Unlike most digital circuits that can be efficiently analyzed at the gate level, SRAM included most analog/mixed-signal circuits must be simulated at the transistor level. Although the IS methods can decrease the number of simulations, the time for a single simulation is still large.

Meta-Modeling: To further speed up the yield estimation, many works [13], [16], [17], [31], [32] try to construct a metamodel to take place of the expensive transistor-level simulations by mapping the process variation into the circuit performance metrics. The works [13], [16], and [13] leverage Polynomial Regression, Gaussian Process, Radial Basis Network respectively, which show good accuracy in the low dimensional scenario. However, these meta models suffer two challenges.

(I) **High Dimensionality:** Most of works [8], [10], [16] only consider the effect of threshold voltage V_{th0}, while other process variations also have significant effects on the SRAM performance, such as offset voltage V_{off} [1] and electron mobility u_0. These parameters increase the unknown coefficients in the model construction. Fig. 1 shows the simulation results of SRAM read access delay with different variations near the failure region. 1000 samples are sorted by their corresponding read performance. There is a huge difference between the simulation results of samples that only consider V_{th0} and the ones that consider all parameters. Furthermore, the number of transistor-level simulations required by the meta models [13], [16], [17] grows exponentially with the total number of process parameters to obtain acceptable accuracy. For example, there are total 4608 process parameters in a 256-depth 6T SRAM column only considering the variations of V_{th0}, V_{off}, u_0 for each transistor. Other state-of-art works, such as [31] and [32], construct the computational efficient meta models by utilizing the sparsity of the underly problem. However, the result of yield estimation is very sensitive to the accuracy of the metamodel, which still needs an extremely large number of training sets to make the model converge.

(II) **Discontinuity:** For the SRAM yield estimation, the model should guarantee the accuracy of the high sigma region of SRAM performance. Thus, a wide range of process parameter variation needs to be considered in the performance meta-model modeling. However, the variations in a certain direction (e.g. positively increased V_{off}) will increase the read access delay and even make the simulation failed which means the simulator can't read a value in these cases. In Fig. 1, the 1000 samples are sorted by their read access delay. And we found the simulation failed at about the 900^{th} sample due to the too large process variation. Notice the maximum continuous result is 5.8×10^{-8}. And to facilitate our statistics, we set the results of these simulation failed samples as a constant, such as "1×10^{-7}" in Fig. 1. The constant can be set to other values larger than the maximum continuous result. This phenomenon comes from two reasons. First, the reliability of bit cells is no longer guaranteed in low supply voltages, resulting in extreme cases where the data stored in cells cannot be read out correctly due to the large process variation. Second, the transient analysis time is limited in the. **TRAN** statement. It is unrealistic to set too large analysis time in the expensive simulation. Oppositely, the time limit is set to the Wordline enable time in the practical SRAM design. The discontinuity region brings a huge challenge in performance modeling. Most previous works [30]–[32] try to construct a high dimensional model to approximate the circuit performance metric by decomposing $f(x)$ into a combination of N orthonormal basis functions. Once the basic functions are determined, the modeling problem is transferred to solve the model coefficients. The termination of such algorithms is to find a set of coefficients to minimize the loss function, typically the Mean Square Error (MSE). However, these regression-based models can only fit the continuous performance because of their nature of approaching all training samples to minimize MSE. It is infeasible to apply these models in the high dimensional and discontinuous scenario simultaneously in SRAM read performance modeling.

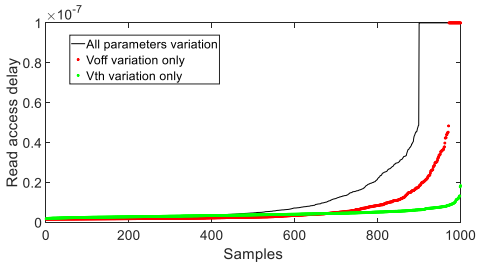


FIGURE 1. The effects of different variation on 0.6 V SRAM read access delay.

In this paper, a statistical SRAM performance meta model is constructed and applied to the SRAM yield estimation. Our paper contributions are summarized as follows:

- We developed a Spline-High Dimensional Model Representation (SP-HDMR) to replace expensive transistor-level simulations in yield estimation. The model is based on Sobol's theory [21] which can decompose a high dimensional problem into multiple low dimensional ones. Meanwhile, a strategy of adaptive modeling with sparsity analysis is developed to

[1] compensate threshold voltage shifting in BSIM model. $Ids = uCox * \exp(\frac{Vgs-vth-voff}{n}) * \exp(Vds/n)$

further minimize the number of unknown coefficients of SP-HDMR in high dimensional scenarios. The spline function is chosen as the kernel of SP-HDMR and properly trained to address the discontinuity problem caused by the large variations of process parameters within limited analysis time.
- A yield analysis framework is developed by integrating our SP-HDMR model into a state-of-art importance sampling method to replace expensive transistor-level simulations. It aggressively improves the yield estimation overhead compared to both the traditional MC method and the pure IS-based method. The accuracy can be guaranteed by the technique of re-simulation.

The rest of this paper is organized as follows. In Section II, the rare event analysis problem and related works are revisited. The yield analysis framework based on SP-HDMR is introduced in Section III. Section IV provides details about the construction of SP-HDMR. The accuracy and efficiency of our method will be demonstrated by several experiments in Section V. In Section 6, we will give our conclusion finally.

II. PRELIMINARIES

A. RARE EVENT ANALYSIS

For the n process variables: $P = [p_1, p_2, \ldots, p_n]$, these variables are modeled as a vector of independent Gaussian variables: $\boldsymbol{x} = [x_1, x_2, \ldots, x_m]$ by the Principle Component Analysis (PCA) [26] in commercial Process Design Kit (PDK). For generalization, each variable is normalized to standard Normal. And $H(\boldsymbol{x})$ is joint probability density function (PDF) of $\boldsymbol{x}$. Let $f(\boldsymbol{x})$ be the interest performance metric which is measured through expensive transistor-level simulation, such as SRAM read access delay in our work.

For the failure rate evaluation of SRAM, we denote S as the tiny failure region. We define the circuit performance doesn't meet the specification when $f(\boldsymbol{x}) \in S$. And we further introduce indicator function $I(\boldsymbol{x})$ to identify pass/fail of $f(\boldsymbol{x})$:

$$I(\boldsymbol{x}) = \begin{cases} 0, & if\ f(\boldsymbol{x}) \notin S \\ 1, & if\ f(\boldsymbol{x}) \in S \end{cases} \tag{1}$$

Therefore, the probability can be calculated as:

$$P_{fail} = P(f(\boldsymbol{x}) \in S) = \int I(\boldsymbol{x}) \cdot H(\boldsymbol{x}) d\boldsymbol{x} \tag{2}$$

Unfortunately, formulation (2) is difficult to calculate analytically because we don't know what distribution $I(\boldsymbol{x})$ satisfies exactly. Traditionally, Monte Carlo is used to estimating the failure probability by sampling from $H(\boldsymbol{x})$ directly, and the unbiased estimate of P_{fail}:

$$\widehat{P}_{fail} = \widehat{P}(Y \in S) = \frac{1}{N} \sum_{i=1}^{N} I(x_i) \overset{N \to +\infty}{\longrightarrow} P(f(\boldsymbol{x}) \in S) \tag{3}$$

B. HIGH DIMENSIONAL IMPORTANCE SAMPLING

For SRAM failure rate estimation, $Y \in S$ is a rare event. Standard MC needs hundred millions of expensive circuit simulations to capture such a "rare event". Although MC can be run at paralleled mode, it is still a time-consuming process.

The IS methods are proposed to reduce the number of simulations by constructing a "distorted" PDF $G(\boldsymbol{x})$ to generate the samples near the failure region. And the failure probability can be expressed as (5):

$$P_{fail} = P(f(\boldsymbol{x}) \in S) = \int I(\boldsymbol{x}) \cdot \frac{H(\boldsymbol{x})}{G(\boldsymbol{x})} \cdot G(\boldsymbol{x}) dx \tag{4}$$

$$= \int I(\boldsymbol{x}) \cdot w(\boldsymbol{x}) \cdot G(\boldsymbol{x}) d\boldsymbol{x} \tag{5}$$

where the $w(\boldsymbol{x})$ denotes the likelihood ratio between original PDF $H(\boldsymbol{x})$ and the distort PDF $G(\boldsymbol{x})$ which compensates for the discrepancy between $H(\boldsymbol{x})$ and $G(\boldsymbol{x})$. And an unbiased IS estimator $\widehat{P}_{IS,fail}$ can be calculated as (6):

$$\widehat{P}_{IS,fail} = \widehat{P}_{IS}(f(\boldsymbol{x}) \in S) = \frac{1}{M} \sum_{k=1}^{M} w(x_k) I(x_k) \overset{a.s.\ M \to +\infty}{\longrightarrow} P(f(\boldsymbol{x}) \in S) \tag{6}$$

With a proper $G(\boldsymbol{x})$, $\widehat{P}_{IS,fail}$ can be approximately equal to MC results. However, the likelihood ratio $w(x_k)$ shows huge numerical instability in the high dimensional scenario [18], where some $w(x_k)$ become dominant and even infinite so that the estimation in equation (6) becomes unreliable.

Wu, etc [9] proposed a provably bounded failure analysis method, High Dimensional Importance Sampling (HDIS). The basic idea of HDIS is to set a new threshold t, where $t > t_c$. And $f(\boldsymbol{x}) > t$ is not a "rare" event but dominates the "rare event" $f(\boldsymbol{x}) > t_c$. Hence the failure rate of SRAM can be estimated as follows:

$$P_{fail}(f(\boldsymbol{x}) > t_c) = P(f(\boldsymbol{x}) > t) \cdot P(f(\boldsymbol{x}) > t_c | f(\boldsymbol{x}) > t) \tag{7}$$

The $P(f(\boldsymbol{x}) > t)$ can be calculated by MC. While the $P(f(\boldsymbol{x}) > t_c | f(\boldsymbol{x}) > t)$ is less than 1 according to conditional probability theory [37], which avoid the huge numerical instability of $P(f(\boldsymbol{x}) > t_c)$ in the high dimension. However, it still needs large number of expensive transistor-level simulations to converge a stable result.

C. META-MODELING

Although IS methods reduce the number of sampling to a certain degree, it is still time-consuming for the expensive simulation overhead, especially the circuit size is large. Just using IS is not enough to decrease the estimation cost drastically compared to parallel MC.

Sobol [21] proposed the High Dimensional Model Representation (HDMR) by decomposing an integrable function into the sum of low dimensional ones, which improves the computational efficiency greatly. It can be formulated as:

$$f(\boldsymbol{x}) = f_0 + \sum_{i=1}^{n} f_i(x_i) + \sum_{1 \le i < j \le n} f_{ij}(x_i, x_j) + \ldots + \sum_{1 \le k_1 < \ldots < k_l \le n} f_{k_{1 \ldots k_l}}(x_{k_1}, x_{k_2}, \ldots, x_{k_l}) + \ldots + f_{1 \ldots n}(x_1, x_2, \ldots, x_n) \tag{8}$$

where f_0 is a constant used to measure the zeroth-order effect of variable vector $\boldsymbol{x}$ on the circuit response $f(\boldsymbol{x})$ and $f_i(x_i)$ represents the first-order effect of a single variable x_i acting independently upon $f(\boldsymbol{x})$. Similarly, $f_{ij}(x_i, x_j)$ represents the second-order effect of variables x_i and x_j on $f(\boldsymbol{x})$. And the latter terms show the high-order effects of the variables. The HDMR expansion aims to represent multivariate functions arising in physical contexts rather than for arbitrary function interpolation [19]. It gives us space to carefully analyze the effects of different variables and train the proper basic functions to characterize these effects. The accuracy and convergence speed of HDMR is determined by its basic functions and corresponding modeling method. Commonly used basic functions, such as Polynomial [24] and Gaussian Process [36], are not accurate enough to handle the discontinuity characterization of SRAM read access delay.

III. YIELD ESTIMATION BASED ON SP-HDMR MODEL

The framework of the proposed yield analysis method, named HDMRIS, is summed as Algorithm 1. We integrate our SP-HDMR model into HDIS [9] to further speedup the yield estimation. The initialization step and the failure calculation ways are the content of HDIS [9] mentioned in the background. The pre-training samples generated by the distorted probability distribution in procedure 2.1 are used to determine the cut point $\boldsymbol{x_0}$ and the training scope of variables in Algorithm 1. After the construction of SP-HDMR, most samples in step 3 are predicted by the model.

Algorithm 1 HDMRIS Overview

Input: Random variables $\boldsymbol{x}$ with its original distribution $H(\boldsymbol{x})$ and the failure specification $f(\boldsymbol{x}) > t_c$

Output: The SRAM failure rate $\widehat{P}_{fail}$

1 Initialization

1.1 Determine the threshold t and calculate the corresponding yield $P_{MC}(f(\boldsymbol{x}) > t)$ by standard MC method.

1.2 Construct the distorted probability distribution $G(\boldsymbol{x})$ by shifting the mean of original distribution $H(\boldsymbol{x})$ and changing the standard deviation of $H(\boldsymbol{x})$.

2 Model Construction

2.1 Generate the pre-training samples from $G(\boldsymbol{x})$ to determine the training scope of variables and cut-point of SP-HDMR.

2.2 Construct SP-HDMR by adaptive modeling method.

3 Failure Rate Calculation

3.1 Generate samples from $G(\boldsymbol{x})$ and evaluate these samples using SP-HDMR.

Re-simulate

3.2 the samples whose predictions are in the predefined range of the failure specification.

3.3 Evaluate the conditional probability $P(f(\boldsymbol{x}) > t_c | f(\boldsymbol{x}) > t)$

3.4 Calculate the failure rate as equation (7)

Notice that transistor-level simulations are only used to evaluate the samples whose predictions are in the predefined range of the failure specification in step 3.2. It is because IS has strict accuracy requirements on the boundary of the failure region. To illustrate it, we review the failure rate calculation in importance sampling shown in (6). As long as the model prediction is wrongly greater than the failure specification, $I(\boldsymbol{x})$ will be incorrectly judged as 1. It makes an unnecessary likelihood $w(x_i)$ in (6) accumulating in $\widehat{P}_{IS,fail}$, resulting in a large error in the final estimation.

As the black and red lines are shown in Fig. 2, the convergence results of yield estimation using transistor-level simulations (HSPICE results) and SP-HDMR only are completely different. As the blue and green lines are shown in Fig. 2, the estimation with the re-simulation technique can greatly reduce the prediction error. Re-simulation within a 7% range of failure specification performs best in this comparison. In our experiment, there are only 532 reevaluated samples in total, which is a small fraction of the whole sample set.

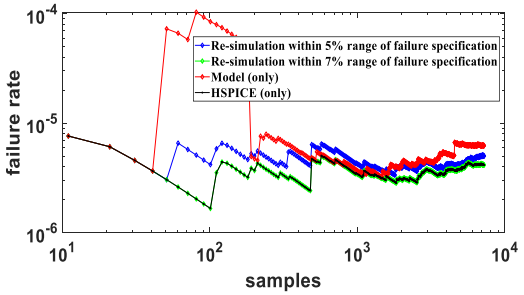


FIGURE 2. Read time failure rate estimation in different modes on the 0.6 v 128 bits SRAM column.

The range of re-simulation depends on the accuracy requirement of the SP-HDMR. The more accurate the meta-model is, the smaller range can be set to satisfy the requirement of importance sampling.

IV. SPLINE HIGH DIMENSIONAL MODEL REPRESENTATION

A. ALGORITHM OVERVIEW

In most well-defined physical systems, only relatively low-order correlations among input variables have significant impacts on the output [19]. Besides, the process variation variables $\boldsymbol{x} = [x_1, x_2, \ldots, x_n]$ has been modeled as independent ones by principal component analysis (PCA) [26] in PDK. Hence, we reserve the top two order terms in (8) to achieve the balance between the complexity and the accuracy. As shown in Table 1, there is almost no difference between the results predicted by the top two order terms and the top three order terms. However, the training cost grows drastically with the exponentially increased second-order interacting terms.

And the proposed SP-HDMR can be formulated as:

$$f(\boldsymbol{x}) = f_0 + \sum_{i=1}^{n} f_i(x_i) \quad s.t.\ \boldsymbol{x} \in (\bar{\boldsymbol{x}} - 8\sigma, \bar{\boldsymbol{x}} + 8\sigma) \tag{9}$$

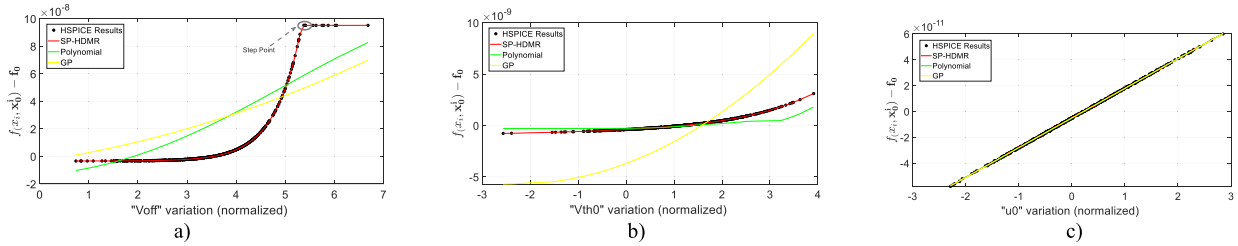


FIGURE 3. The effects of different process variables on Read access delay under 0.7 v 40 nm SRAM cell. Here, Δ (*Read access delay*) = $f(x_i, x_0^i) - f_0$ represents the effects of different variables and "HSPICE results" represents the circuit simulations. a) The effect of "V_{off}." b) The effect of "V_{th0}." c) The effect of "u_0."

TABLE 1. Relative error *v.s.* training cost with different orders on the 0.6 V SRAM cell read access delay.

Num of orders	Relative Err.	Training Samples
Top two	1.3%	232
Top three	1.1%	3400

where the $\bar{x}$ is the mean of variables and $f(\boldsymbol{x})$ denotes SRAM read access delay. The variation range of $\boldsymbol{x}$ is set to $\pm 8\sigma$ which has the probability of 10^{-15}, which means the samples out of this range are infeasible.

There is no unique expansion for equation (9) [19]. However, the modeling cost heavily depends on its expansion. Cut-HDMR expansion [19] provides an exact representation of equation (9) along a hyperplane passing through the "cut" point or the reference point. Then $f(\boldsymbol{x})$ is represented by superposing the value of $f(\boldsymbol{x})$ on the lines, planes, and hyper-planes passing through the cut point. As a result, each term of (9) can be modeled as:

$$\begin{cases} f_0 = f(\boldsymbol{x_0}) \\ f_i(x_i) = f\left(x_i, \boldsymbol{x_0^i}\right) - f_0 \end{cases} \tag{10}$$

where $\boldsymbol{x_0} = [x_{10}, x_{20}, \ldots, x_{n0}]^T$ is the cut point, the $\boldsymbol{x_0^i} = [x_{10}, x_{20}, \ldots, x_{(i-1)0}, x_{(i+1)0}, \ldots, x_{n0}]^T$ represents the variable vector $\boldsymbol{x_0}$ without the element x_{i0}.

The cut-HDMR expansion only involves simple arithmetic computation, which provides us the flexibility of selecting basic functions. In this work, spline basic functions and adaptive training method with sparsity analysis are developed to minimize cost, which is discussed in the next subsection.

B. IMPLEMENTATION DETAILS

1) BASIC FUNCTION SELECTION

The key to constructing an accurate HDMR is to obtain the proper basic functions for measuring each order effect accurately. Due to the too large process variations in the failure region, the simulations often failed to get numerical values. Here, we define SRAM read access failure as that the read access delay exceeds the 4.8×10^{-8} seconds (4.5σ) at 0.7V. The word "failed" will be written into the final data file. For our modeling convenience, these discontinuous responses are represented by a constant. However, it brings a "jump" in the trend of performance metric, which is a stepping point as marked in Fig. 3 a). The characteristic of regression functions, such as Polynomial [24] and Gaussian process [32], is to reach all data points as close as possible for minimizing some cost (e.g., the square of error), which make them failed to fit the discontinuous curves.

To address this issue, we notice that the interpolation method can pass through all samples. It can model the discontinuous functions with proper intervals. And among the different interpolation functions, the spline function $p(x), x \in [a, b]$ is the most widely used interpolant in the spline interpolation for its smoothness and robustness (High order interpolation will lead to the "Runge phenomenon" [38]). It is defined by piecewise third-order polynomials defined as follows:

$$p(x_i) = \begin{cases} p_1(x_{i1})\, a < x_{i1} < x_i^{(1)} \\ p_2(x_{i2})\, x_i^{(1)} < x_{i2} < x_i^{(2)} \\ \vdots \\ p_k(x_{ik})\, x_i^{(k-1)} < x_{ik} < x_i^{(k)} \\ p_{k+1}(x_{ik+1})\, x_i^{(k)} < x_{ik+1} < x_i^{(k-dis)} \\ \vdots \\ p_n(x_{in})\, x_i^{(n-1)} < x_{in} < b \end{cases}$$
$$s.t.\ i = 1, \ldots m; k = 1, \ldots, n \tag{11}$$

where $p(x_i)$ is the spline basic function of ith variable, x_i. The $x_i^{(k)}$ is the split point and $x_i^{(k-dis)}$ is the first split point in the discontinuous region. And $p_k(x_{ik})$ is the third-order polynomials for the k-th interval. The $p(x)$ can model discontinuous functions as long as there is one split point fallen in the discontinuous region. Notice that the first split point "$x_i^{(dis-k)}$" in equation (11) affects the shape of $p(x)$. The $x_i^{(dis-k)}$ should be close to the step point. If the "$x_i^{(dis-k)}$" is away from the step point, the $p(x_i)$ will have a relatively large error near the step point. Hence the training samples must be carefully generated in our adaptive modeling method, which is discussed in the next subsection.

Fig. 3 compares the fitting effects of different basic functions for three process parameter variables, offset voltage V_{off}, threshold voltage V_{th0}, and electron mobility u_0, by using 50 training samples near the failure boundary. The "$f(x_i, \boldsymbol{x_0^i}) - f_0$" in y-axis means the first-order effects on SRAM read access delay of these three variables. 100 sample predictions are sorted by the normalized variable. As shown in Fig. 3 (a), the SRAM read access time failed to be evaluated

by transistor-level simulation in the tail of V_{off}. The variable V_{off} affects the drain-source current of MOSFET [27]. Too large variation on V_{off} may make the low gate-source voltage and cannot open the channel between drain and source so that $I_{ds} \approx 0$. It is the reason that why the simulation failed to get an exact SRAM read access delay. As the result, all basic functions failed to predict the correct effects except for the spline function. Notice that, the trend of two functions, Gaussian Process and Polynomial, is like a "balance" process. However, the discontinuous results make these regression functions "balance" wrongly to average out the overall error. Oppositely, the spline function is a kind of interpolation and jumps to discontinuous value "1×10^{-7}" successfully because it must pass through all training samples. As shown in Fig. 3 b), the spline function also outperforms other methods in measuring the effect of V_{th0} on the performance within 50 training samples, although there is strong nonlinearity between V_{th} and SRAM read access delay. All functions can fit the linear effects of u_0 perfectly as shown in Fig. 3 c).

Notice that the weighted regression is also useful to fit the curve by penalizing more the deviation near the discontinuous region. The goal of weighted regression is to ensure that each data point has an appropriate level of influence on the parameter estimates. It requires that we know exactly what the weights are. However, the optimal weights, which are based on the true variances of each data point, are never known. Estimated weights have to be used instead. The effect of using estimated weights is difficult to access. When the weights are estimated from a small number of training samples, the results of an analysis can be very badly and unpredictably affected. As shown in Fig. 4, the weight regression with 200 training samples can fit the discontinuous region accurately. Whereas the spline just takes 50 training samples. Notice that there is no need to consider the model sensitivity to noise because we only estimate the deterministic effects of process parameters on the circuit responses without injecting any noise in this work.

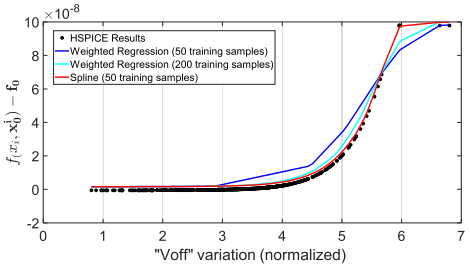


FIGURE 4. Fitting curve of Weighted Regression on V_{off} with different training cost.

2) ADAPTIVE MODELING METHOD

In order to reduce the training cost, we developed an adaptive sampling strategy with sparsity analysis to support the proposed model. Compared with the traditional method of collecting a large number or even all of the required samples at a time by a certain algorithm, the proposed sampling strategy collects only one or several samples at a time with respect to the different impacts of variables on the performance metric. It can be viewed as the local sensitivity analysis for these variables before constructing basic functions for them. SP-HDMR's training method is as follows:

Algorithm 2 Adaptive Modeling Method

/∗Cut point selection∗/

1: Choose a proper "cut" center $\boldsymbol{x_0} = [\boldsymbol{x_{10}}, \boldsymbol{x_{20}}, \ldots, \boldsymbol{x_{n0}}]^T$ and get the corresponding response $\boldsymbol{f_0}$ by running simulation.

/∗Basic functions training with sparsity analysis∗/

2: Modeling the first-order basic function $\boldsymbol{f}(\boldsymbol{x_i})$ by linear interpolation at the two points collected in the close range of the lower bound and upper bound of $\boldsymbol{x_i}$.

3: Check the linearity of $\boldsymbol{f_i}(\boldsymbol{x_i})$. If $\boldsymbol{f}(\boldsymbol{x_i})$ goes through $\boldsymbol{x_0}$, $\boldsymbol{f}(\boldsymbol{x_i})$ is considered as linear and the construction of $\boldsymbol{f}(\boldsymbol{x_i})$ terminates. Otherwise, $\boldsymbol{f}(\boldsymbol{x_i})$ should be reconstructed by a spline function in step 4.

4: Use the Latin Hypercube sampling to construct $\boldsymbol{f}(\boldsymbol{x_i})$

- 4.1 Set the initial interval as 10.
- 4.2 Construct the spline function to measure this effect.
- 4.3 Compare the simulation results with the function predictions. If the average error is large than the threshold, increase the interval and generate new samples. Then, go to step 4.2 to reconstruct $\boldsymbol{f}(\boldsymbol{x_i})$.

/∗Complete the SP-DHRM∗/

5: Repeat steps 2 $\sim$ 4 until all the first-order component functions are obtained.

In step 1, the choice of the "cut" point $\boldsymbol{x_0} = [x_{10}, x_{20}, \ldots, x_{n0}]^T$ can be random if the equation (10) is taken out to convergence. Given that it is important to fit the tail of the delay distribution for estimating SRAM failure rate accurately. The cut point can be chosen as the mean of samples near the failure region.

We find that the majority of variables only have weak effects on SRAM read access delay, which can be viewed as a sparsity constraint on SRAM performance modeling. To facility this sparsity, we execute the sparsity analysis for each variable in step 2 and step 3. If the slope of $f(x_i)$ in step 3 is zero, it means the variable x_i does not affect SRAM read access delay and can be filtered out in the modeling. In another scenario, if $f(x_i)$ goes through x_0, it means a linear function is sufficient to model the relationship between variable x_i and the target performance metric. Otherwise, $f(x_i)$ will be reconstructed by the spline basic function.

To characterize the discontinuous effects of variables, the training samples must cover a relatively wide range to cross continuous and discontinuous regions. Hence, we adopt the Latin-hypercube sampling [35] to generate the initial samples across different intervals of the cumulative probability density for each variable in step 4. For the algorithm efficiency, we start sampling from 10 intervals empirically and increase the intervals until finding the split point "$x_i^{(k-dis)}$" closest the step point.

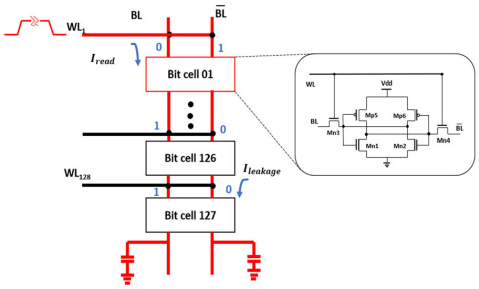


FIGURE 5. The schematic of SRAM column. Inset: Schematic of 6T SRAM cell.

V. EXPERIMENTAL RESULT AND COMPARISON

A. EXPERIMENT SETUP

The proposed SRAM performance metamodel will be first verified on the SRAM column and sense amplifier by comparing with SPICE simulations. We also implement other state-of-art meta-models, such as LRTA [32], and OMP [31] for accuracy comparison under different operating voltages. And then the yield analysis based on our model will be illustrated by comparing with Monte Carlo which runs in parallel mode on 60 cores server and the other two importance sampling methods, HDIS [9] and AIS [6]. All experiments are performed with 40nm SMIC model on the Server with Intel Xeon Gold 5118 CPU @ 2.30 GHz.

B. EXPERIMENT ON SRAM COLUMN

1) MODEL ACCURACY VALIDATION

Fig. 5 shows the simplified schematic of the read path of the 128-row SRAM column which has 2306 variables. The read operation begins by activating word-line (WL) and the pre-charged bit-lines. One bit-line *BL* will discharge through the first accessed cell and enlarges the voltage difference between *BL* and $\overline{BL}$. The read access delay is defined as the time required to generate the voltage difference between two bit-lines that can be sensed by the sense amplifier. Notice that to generate the worst case for the read operation, the accessed *Bit cell 1* stores "0" and other idle cells store "1", which maximum the leakage current through idle bits to increase the read access delay and impede the successful read. For the SRAM yield estimation, we consider the read access delay failure that the time of read operation exceeds a specified time.

To verify the accuracy and efficiency of our model and modeling method, we trained other state-of-art high dimensional models, LRTA [32] and OMP [31], with 3000 training samples near the failure region.

As shown in Fig. 6, a hundred samples near the failure region are predicted by meta-models and transistor-level simulations, respectively. All samples are sorted by the read delay measured by transistor-level simulations. Fig. 7 summarized the average error of three models from Fig. 6. In the Fig. 6 a) and b), the SRAM performance shows strong nonlinearity, all models can fit the relationship with enough accuracy at 0.8V and 0.9V. The average error of SP-HDMR, OMP and LRTA at 0.8V is 2.1%, 8.8%, 4.3% respectively. And the relative error of all models at 0.9V is even lower than 2%. Fig. 6 c) and Fig. 6 d) show the predicted read delay at 0.7V and 0.6V, respectively. The SRAM read access delay shows varying degrees of discontinuity. The predicted values of SP-HDMR are closest to the results of simulations. However, as shown in Fig. 6 c) and d), OMP and LRTA have the deviation from the HSPICE results in the first half and the second half of the curve. The relative error of OMP and LRTA at 0.6V has reached 29.6% and 14.8% respectively,

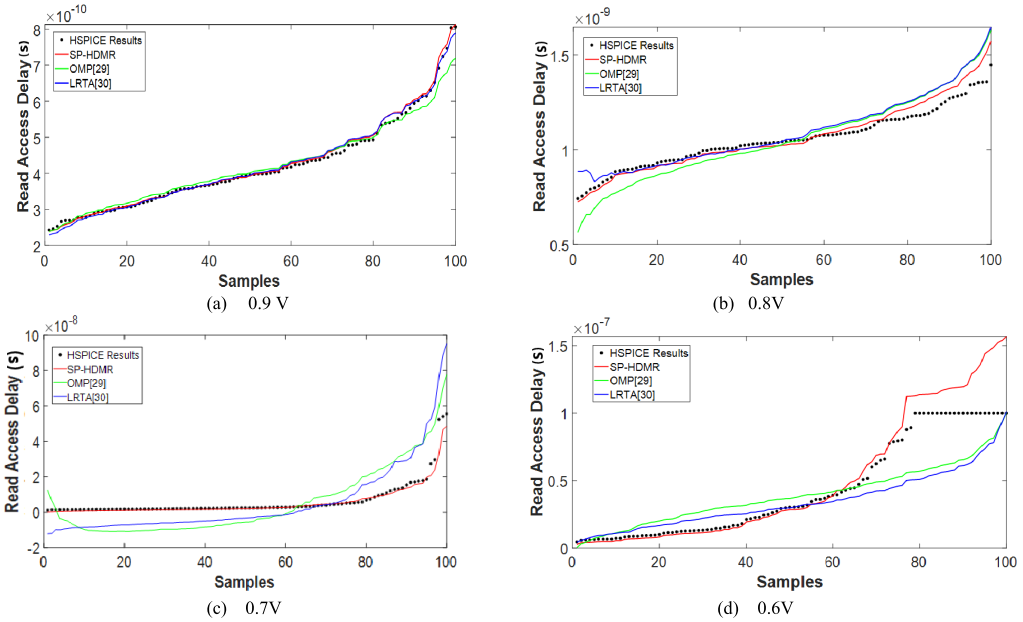


FIGURE 6. The fitting curves of different models under the different voltages on the 128 bits SRAM column.

TABLE 2. Accuracy and efficiency comparison on 2306 dimensional 40 nm 0.6 v 128 bits SRAM column (the time per simulation on our platform is 0.09 seconds while the time of model prediction is just 0.007 milliseconds).

	MC	AIS	HDIS	HDMRIS
Failure Prob.	1.45e-5 (0%)	2.53e-4 (failed)	1.40e-5(3.4%)	1.58e-5 (8.9%)
Pre-sampling (#sim)	0	4000	8000	8000
Importance Sampling (#sim)	10e7	12000	65820	10052
Importance Sampling (#model)	0	0	0	58948
Total (#sim)	66.9h	-	3.55h (18.8x)	0.67h(99.9.x)

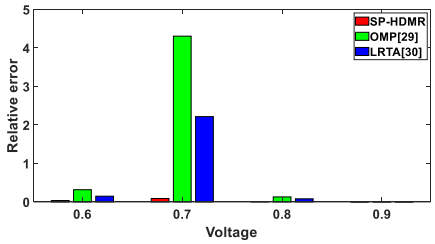


FIGURE 7. The relative error of different methods corresponding to Fig. 6.

while the SP-HDMR is just 4.6%. The huge error of OMP and LRTA in modeling read access delay can be attributed to the natural characteristic of regression of attempting to be as close to all real values as possible to minimize some cost, usually the mean of squares of errors. However, the discontinuous values break the normal trend of performance, which brings a larger error for the regression models. If we set the "failed" result as "1×10^{-4}", the error of predictions of OMP and LRTA will be enlarged dramatically due to this characteristic. Notice that the relative error of OMP and LRTA at 0.7V is much larger than that of those at 0.6V in Fig. 6. It is because that the too large gap between the successful simulated results, the magnitude of 10^{-9}, and the set failed results, the magnitude of 10^{-7} at 0.7V.

2) YIELD ANALYSIS EFFICIENCY VALIDATION

We compared the accuracy and efficiency of the failure rate estimated by our method with MC, AIS, and the original HDIS without modification.

We apply Fig of Merit (FOM) ρ to verify our proposed method, defined in [10]:

$$\rho = \frac{\sqrt{VAR_{\widehat{P}_{fail}}}}{\widehat{p}_{fail}} \tag{12}$$

where the $VAR_{\widehat{P}_{fail}}$ is the variance of $\widehat{P}_{fail}$. And $\rho < \varepsilon\sqrt{log(1/\delta)}$ means one estimation has reached $(1-\varepsilon)\,100\%$ accuracy with $(1-\delta)\,100\%$ confidence. Here, we set $\rho = 0.1$ which means 90% accuracy with a 90% confidence level. As shown in Fig. 8, the AIS has failed to converge the right result when the FOM reaches 0.1. It is because the diversity of samples is decreasing as the iterations of the resampling procedure in the high dimensional scenario. While HDIS and HDMRIS successfully converge to the required precision with 3.4% error and 8.9% error respectively.

The efficiency of MC, AIS, HDIS, HDMRIS is shown in Table 2. In this experiment, 2160 training samples are included in the importance sampling step of HDMRIS. Both HDIS and HDMRIS cost over 60000 samples to get stable results. Although the relative error of HDMRIS is larger than HDIS, it only consumes 0.67 hours, which is 5.3x faster than HDIS and 99.9x faster than MC.

Besides, we also present HDMRIS under different supply voltages, as shown in Fig. 9. Our method can predict yield accurately for a wide range of supply voltage. The yield estimation results deviate from the simulations when VDD is larger than 0.68V. This is because the relative error of SP-HDMR is enlarged so that the 7% range of re-stimulation

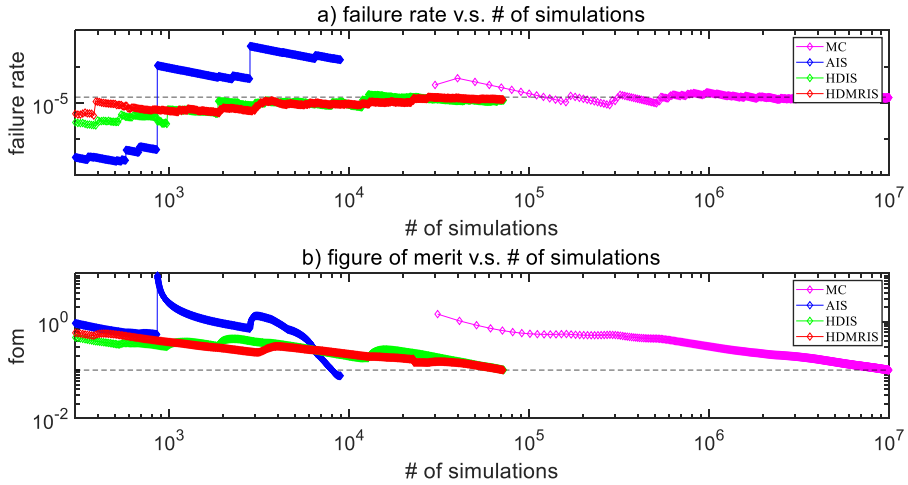


FIGURE 8. Evolution Comparison of Failure Prob. and FoM on 128 bits SRAM column.

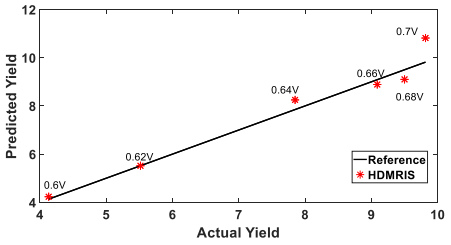


FIGURE 9. The yield prediction of HDMRIS under different voltage on SRAM read path. The reference is simulated by HSPICE.

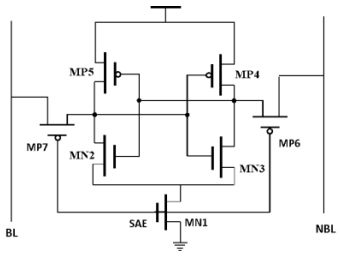


FIGURE 10. Schematic of sense amplifier.

is not enough to ensure an extremely accurate result. We can increase the pre-defined range of re-simulation to reduce the error at the expense of additional simulations.

C. EXPERIMENT ON SENSE AMPLIFIER

We also make an experiment on Sense Amplifier (SA) which is one of the most important components in SRAM circuits. As shown in Fig. 10, it mainly composes of two cross-coupled inverters. The voltage difference between BL and NBL will be enlarged to the expected value due to the positive feedback. We focus on the estimation of offset voltage which is the most important performance metric of SA. It is the voltage difference ΔV between SA inputs (bit-lines) when the inverters of SA remain at the metastable point. The larger offset the more time will be needed for SRAM to discharge one of the bit-lines, which leads to additional read delay. We consider the failure of SA that the offset voltage exceeds 28 mV due to process variation.

Fig. 11 shows 100 offset voltage predictions at different voltages. The predicted results under different operating voltages are less different. It is because the SA has less sensitivity on process parameters due to its large transistor size. All models have good accuracy on the SA within the relative error of 5%. Fig. 12 shows the yield analysis result at 0.6V. The MC result remains "golden standard" and converges to 3.13e-5 with 26 hours. While the AIS, HDIS, and HDMRIS take 15 minutes, 36.7 minutes, 10.9 minutes, to converge to 3.21e-5, 3.19e-5, 3.2e-5, respectively. Notice that the time cost of HDMRIS includes 6957 simulations and 18640 model predictions. NOTICE 960 simulations are used to train the SP-HDMR model. AIS and HDIS take 9600 simulations and 23400 simulations respectively. All methods have converged to a reasonable result with less than 4% relative error when

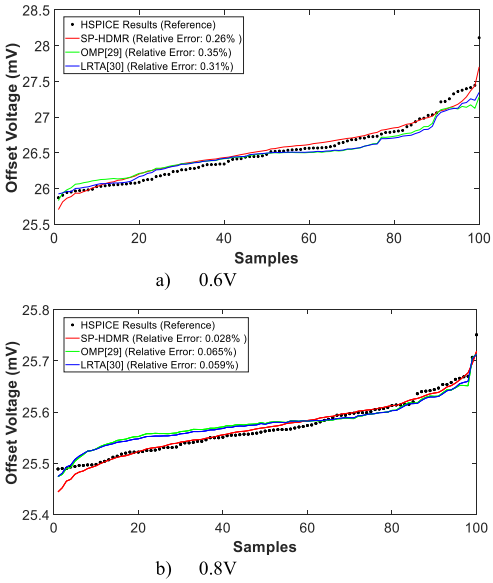


FIGURE 11. The fitting curves of different models under the different voltages on the sense amplifier.

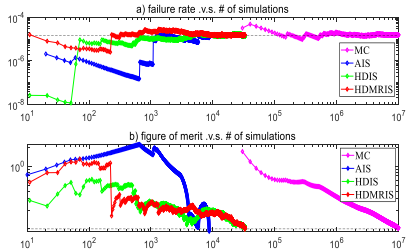


FIGURE 12. Evolution Comparison of Failure Prob. and FoM on Sense Amplifier.

their FoMs are lower 0.1. However, our method speedup 1.3X over AIS and 3.4X over HDIS, which gains 143.1X over MC.

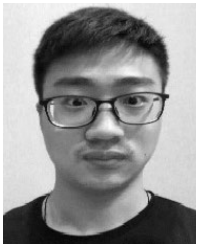
VI. CONCLUSION

In this paper, we developed a yield analysis framework HDMRIS based on our SP-HDMR model to accelerate yield analysis by substituting expensive transistor-level simulations. To construct SP-HDMR model, we facility the cut-HDMR expansion to provide a computationally efficient model representation. Then the spline basic function is carefully analyzed and trained to address the discontinuous problem brought by the large variations of process parameters in SRAM. And the model is implemented efficiently by an adaptive sampling strategy with sparsity analysis. The proposed HDMRIS achieves great speedup compared to the state-of-the-art yield estimation methods with enough accuracy.

REFERENCES

[1] A. J. Bhavnagarwala, X. Tang, and J. D. Meindl, "The impact of intrinsic device fluctuations on CMOS SRAM cell stability," *IEEE J. Solid-State Circuits*, vol. 36, no. 4, pp. 658–665, Apr. 2001.
[2] A. P. Chandrakasan, W. J. Bowhill, and F. Fox, *Design of High-Performance Microprocessor Circuits*. Hoboken, NJ, USA: Wiley, 2000.
[3] J. C. Maxwell, *A Treatise on Electricity and Magnetism*, vol. 2, 3rd ed. Oxford, U.K.: Clarendon, 1892, pp. 68–73.

[4] S. Shen, T. Shao, X. Shang, Y. Guo, M. Ling, J. Yang, and L. Shi, "TS cache: A fast cache with timing-speculation mechanism under low supply voltages," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 28, no. 1, pp. 252–262, Jan. 2020.
[5] R. V. Joshi, S. Mukhopadhyay, D. W. Plass, Y. H. Chan, C.-T. Chuang, and Y. Tan, "Design of sub-90 nm low-power and variation tolerant PD/SOI SRAM cell based on dynamic stability metrics," *IEEE J. Solid-State Circuits*, vol. 44, no. 3, pp. 965–976, Mar. 2009.
[6] X. Shi, F. Liu, J. Yang, and L. He, "A fast and robust failure analysis of memory circuits using adaptive importance sampling method," in *Proc. 55th ACM/ESDA/IEEE Design Autom. Conf. (DAC)*, Jun. 2018, pp. 1–6.
[7] W. Wu, W. Xu, R. Krishnan, Y.-L. Chen, and L. He, "REscope: High-dimensional statistical circuit simulation towards full failure region coverage," in *Proc. 51st ACM/EDAC/IEEE Design Autom. Conf. (DAC)*, Jun. 2014, pp. 1–6.
[8] 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.
[9] W. Wu, F. Gong, G. Chen, and L. He, "A fast and provably bounded failure analysis of memory circuits in high dimensions," in *Proc. 19th Asia South Pacific Design Autom. Conf. (ASP-DAC)*, Jan. 2014, pp. 424–429.
[10] L. Dolecek, M. Qazi, D. Shah, and A. Chandrakasan, "Breaking the simulation Barrier: SRAM evaluation through norm minimization," in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Design*, Nov. 2008, pp. 322–329.
[11] M. Qazi, M. Tikekar, L. Dolecek, D. Shah, and A. Chandrakasan, "Loop flattening & spherical sampling: Highly efficient model reduction techniques for SRAM yield analysis," in *Proc. Des., Automat. Test Eur. Conf. Exhib. (DATE)*, Mar. 2010, pp. 801–806.
[12] W. Wu, S. Bodapati, and L. He, "Hyperspherical clustering and sampling for rare event analysis with multiple failure region coverage," in *Proc. Int. Symp. Phys. Design*, Apr. 2016, pp. 153–160.
[13] M. Wang, C. Yan, X. Li, D. Zhou, and X. Zeng, "High-dimensional and multiple-failure-region importance sampling for SRAM yield analysis," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 25, no. 3, pp. 806–819, Mar. 2017.
[14] K. Katayama, S. Hagiwara, H. Tsutsui, H. Ochi, and T. Sato, "Sequential importance sampling for low-probability and high-dimensional SRAM yield analysis," in *Proc. IEEE/ACM Int. Conf. Computer-Aided Design (ICCAD)*, Nov. 2010, pp. 703–708.
[15] L. Pang, M. Yao, and Y. Chai, "An efficient SRAM yield analysis using scaled-sigma adaptive importance sampling," in *Proc. Design, Autom. Test Eur. Conf. Exhib. (DATE)*, Mar. 2020, pp. 97–102.
[16] M. Wang, W. Lv, F. Yang, C. Yan, W. Cai, D. Zhou, and X. Zeng, "Efficient yield optimization for analog and SRAM circuits via Gaussian process regression and adaptive yield estimation," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 37, no. 10, pp. 1929–1942, Oct. 2018.
[17] J. Yao, Z. Ye, and Y. Wang, "An efficient SRAM yield analysis and optimization method with adaptive online surrogate modeling," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 23, no. 7, pp. 1245–1253, Jul. 2015.
[18] T. B. B. Li and P. Bickel, "Curse-of-dimensionality revisited: Collapse of importance sampling in very high-dimensional systems," Dept. Statist., Univ. California, Berkeley, Berkeley, CA, USA, Tech. Rep. 696, 2005.
[19] H. Rabitz and Ö. F. Alis, "General foundations of high-dimensional model representations," *J. Math. Chem.*, vol. 25, no. 2, pp. 197–233, 1999.
[20] S. Shan and G. G. Wang, "Metamodeling for high dimensional simulation-based design problems," *J. Mech. Design*, vol. 132, no. 5, May 2010, Art. no. 051009.
[21] I. M. Sobol, "Sensitivity estimates for nonlinear mathematical models," *Math. Model. Comput. Exp.*, vol. 1, no. 4, pp. 407–414, 1993.
[22] S. Shen, T. Shao, M. Ling, J. Yang, and L. Shi, "Modeling and designing of a PVT auto-tracking timing-speculative SRAM," in *Proc. Design, Autom. Test Eur. Conf. Exhib. (DATE)*, Mar. 2020, pp. 1073–1078.
[23] J. Yang, H. Ji, Y. Guo, J. Zhu, Y. Zhuang, Z. Li, X. Liu, and L. Shi, "A double sensing scheme with selective bitline voltage regulation for ultralow-voltage timing speculative SRAM," *IEEE J. Solid-State Circuits*, vol. 53, no. 8, pp. 2415–2426, Aug. 2018.
[24] R. Y. Rubinstein and P. W. Glynn, "How to deal with the curse of dimensionality of likelihood ratios in Monte Carlo simulation," *Stochastic Models*, vol. 25, no. 4, pp. 547–568, 2009.
[25] J. R. Edwards, "Alternatives to difference scores: Polynomial regression and response surface methodology," in *Advances in Measurement and Data Analysis*. 2002, pp. 350–400.
[26] I. Miller, "Probability, random variables, and stochastic processes," *Technometrics*, vol. 8, no. 2, pp. 378–380, May 1966.
[27] *HSPICE MOSFET Models Manual, Version D-2010.12*, Synopsys, Mountain View, CA, USA, Dec. 2010.
[28] M. E. Sinangil, H. Mair, and A. P. Chandrakasan, "A 28 nm high-density 6T SRAM with optimized peripheral-assist circuits for operation down to 0.6V," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2011, pp. 260–262.
[29] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Stat. Soc. B, Methodol.*, vol. 58, no. 1, pp. 267–288, Jan. 1996.
[30] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Ann. Statist.*, vol. 32, no. 2, pp. 407–499, 2004.
[31] X. Li, "Finding deterministic solution from underdetermined equation: Large-scale performance variability modeling of analog/RF circuits," *IEEE Trans. Comput.-Aided Design. Integr. Circuits Syst.*, vol. 29, no. 11, pp. 1661–1668, Nov. 2010.
[32] X. Shi, H. Yan, Q. Huang, J. Zhang, L. Shi, and L. He, "Meta-model based high-dimensional yield analysis using low-rank tensor approximation," in *Proc. 56th Annu. Design Autom. Conf.*, Jun. 2019, pp. 1–6.
[33] X. Li and H. Liu, "Statistical regression for efficient high-dimensional modeling of analog and mixed-signal performance variations," in *Proc. 45th Annu. Conf. Design Autom. (DAC)*, 2008, pp. 38–43.
[34] A. Singhee and R. A. Rutenbar, "Beyond low-order statistical response surfaces: Latent variable regression for efficient, highly nonlinear fitting," in *Proc. 44th ACM/IEEE Design Autom. Conf.*, Jun. 2007, pp. 256–261.
[35] X. Li, J. Le, and L. T. Pileggi, "Statistical performance modeling and optimization," *Found. Trends Electron. Design Autom.*, vol. 1, no. 4, pp. 331–480, Sep. 2006.
[36] C. K. I. Williams, and C. E. Rasmussen, *Gaussian Processes for Machine Learning*, vol. 2, no. 3. Cambridge, MA, USA: MIT Press, 2006.
[37] A. Papoulis and S. Pillai, *Probability, Random Variables, and Stochastic Processes*. New York, NY, USA: McGraw-Hill, 2001.
[38] D. Chen, T. Qiao, H. Tan, M. Li, and Y. Zhang, "Solving the problem of Runge phenomenon by pseudoinverse cubic spline," in *Proc. IEEE 17th Int. Conf. Comput. Sci. Eng.*, Dec. 2014, pp. 1226–1231.



LIANG PANG received the B.S. degree in communication engineering from the Hefei University of Technology, Hefei, China, in 2017. He is currently pursuing the Ph.D. degree in microelectronics and solid state electronics with Southeast University, Nanjing, China. His research interests include low-voltage SRAM design and reliability analysis.

SHAN SHEN was born in 1993. He received the B.S. degree from the Department of Microelectronics, Jiangnan University, in 2016. He is currently pursuing the Ph.D. degree in microelectronics with the School of Microelectronics, Southeast University.

His research interests include hardware designs in computer architecture and memory systems.

MENGYUN YAO received the B.S. degree from Southwest Jiaotong University, China, in 2017, and the M.S. degree from Southeast University, Nanjing, China, in 2020. Her research interests include low-voltage SRAM design and reliability optimization.