An Efficient Yield Estimation Method for Layouts of High Dimensional and High Sigma SRAM Arrays

Yue Shen[1], Changhao Yan[1*], Sheng-Guo Wang[2], Dian Zhou[3] and Xuan Zeng[1*]
[1] State Key Lab of ASIC & System, School of Microelectronics, Fudan University, Shanghai, China
[2] Dept. of CS, University of North Carolina at Charlotte, Charlotte, USA
[3] Dept. of EE, University of Texas at Dallas, Dallas, USA

***Abstract*—This paper firstly focuses on yield estimation problem on post-layout-simulation of high dimensional SRAM arrays. Post-layout-simulation is much more credible than pre-simulation. However, it introduces strong relationship among SRAM columns. The Multi-Fidelity Gaussian Process model between the small and the large SRAM arrays near Optimal Shift Vector (OSV) is built. An iterative strategy is proposed and Multi-Modal method is applied to obtain more prior knowledge of the small SRAM arrays and further accelerate convergence. Experimental results show that the proposed method can gain 5-7x speedup with less relative errors than the state-of-the-art method for 384D cases.**

I. Introduction

As fabrication technology scaling down to nanometers, yields of industrial SRAM arrays usually consisting of millions of bit-cells suffer from process fluctuations. Therefore, accurate failure rate estimations for both pre-simulation and post-simulation should be accomplished within acceptable time costs. The main challenges of SRAM yield estimation are the extremely low failure rate (smaller than 10^{-6}) [1] and the high dimensionality [2]. The simulation time of SRAM arrays grows extraordinarily as O(n^3), where *n* is the number of bit-cells [3]. Therefore, substantially reducing samples of large SRAM arrays is the key to estimate their failure rates.

Monte Carlo (MC) method is the most well-known and widely-used method to estimate yield. However, the extremely low failure rate of SRAM circuits requires 10^7 or even 10^8 samples, which is intolerable for large SRAM arrays.

Importance Sampling (IS) method [4] is widely used in SRAM yield estimation. The basic idea of IS is to find the Optimal Shift Vector (OSV), i.e., the failure point nearest to the origin, and sample near this OSV. Many prior works are based on the IS method. Importance Boundary Sampling (IBS) [5] can handle multiple failure regions effectively in low dimensional cases, but it hardly works in high-dimensional cases. Ref. [6] (MFRIS) applies Multiple Starting Points (MSP) and Sequential Quadratic Programming (SQP) process to search OSVs in linear time complexity for high dimensional cases. However, it needs a repetition at each failure region. Ref. [3] applies Bayesian inference to obtain high dimensional yield distributions from low ones and gains 6x speedup to MFRIS. However, the method of finding OSVs is the same as [6], which dominates the number of samples after efficiency improvement.

In industrial productions, post-simulations on layouts are more credible than pre-simulations for designers, since the influences of parasitic parameters of interconnects and transistors are taken into consideration. In this paper, we firstly focus on estimating the yield on layouts of large SRAM arrays, which has not been mentioned in existing papers, as far as we know.

In pre-simulations, with the assumption of the ideal conducting word lines, bit columns are independent indeed, and the complicated influences on the delays of read/write operations among columns can never be observed. In post-simulations, however, experimental results show that the concerned performances, i.e., voltage difference of bit lines of each column, have similar values with strong correlations. Therefore, the performances of multiple SRAM columns are interested and failure rate estimation on post-simulation of an SRAM array will definitely have multiple failure regions.

Multiple failure regions then lead to multiplying growth of samples in OSV searching, which is a crucial step in most IS-based methods. For a single failure region, the time complexity of the state-of-the-art OSV searching algorithm MFRIS is roughly linear $O(n)$ with respect to the dimensionality of circuits *n*. For *m* multiple failure regions, however, it becomes $O(m \times n)$. Therefore, searching OSV dominates the total simulation cost which is unacceptable for practical SRAM arrays. In a word, although the existing algorithms developed for pre-simulation can be easily ported to post-simulation, they will be invalid due to their sharply degenerate efficiencies.

To accelerate the OSV searching for multiple failure regions efficiently, we utilize the similarity between small SRAM arrays and large ones, which is reasonable and can be easily observed from experimental data. Specifically, we intensively analyze the OSVs in small SRAM arrays first, where the time of SPICE simulations is ignorable compared with large cases. Then, we transfer such knowledge to large arrays and reduce simulations on large arrays obviously.

The main contributions of this paper include: (1) A novel and efficient yield estimation method is firstly proposed for layouts of high dimensional and high sigma SRAM arrays. Multi-Fidelity Gaussian Process (MFGP) [7] on performances of columns between small SRAM arrays (low dimensional) and large ones (high dimensional) are constructed. Because both the MFGP models and OSVs are unknown, an adaptive and iterative scheme is applied, starting with the OSVs of small SRAM arrays. (2) A Multi-Modal (MM) optimization problem is further formulated to find the failure boundaries near OSVs. These failure boundaries serve as supplemental knowledge to accelerate the convergence of the iterative flow. (3) Experimental results show that SPICE simulations on the high dimensional circuit can be reduced to 300-400 samples with 2.4%-3.7% relative errors, which are 3-7x speedup compared with the most well-known methods. Meanwhile, the

*Corresponding authors: {yanch, xzeng}@fudan.edu.cn

978-3-9819263-5-4/DATE21/©2021 EDAA

proposed method can still work on pre-simulations, with about 1.7x speedup.

The rest of this paper is organized as follows. In section II, the problem of yield estimation will be formulated, and we will introduce background knowledge of our proposed work. In section III, the proposed method will be explained in detail. In section IV, the experimental results will be presented. In section V, a final conclusion will be drawn briefly.

II. Background

A. Problem formulation

Let $x = [x_1, \ \ldots, x_D]^T$ be a D-dimensional random variable that represents all the random process variables. We suppose these variables are mutually independent and standard normal. Then the joint probability density function (PDF) $p(x)$ can be written as

$$p(x) = \prod_{d=1}^{D} p_d(x_d) \tag{1}$$

where $p_d(x_d)$ is an independent PDF of process parameters given by foundries, usually a normal distribution. The total failure rate of an SRAM array can be represented as

$$P^{fail} = \int_{\Omega} p(x)dx = \int I(x \in \Omega)p(x)dx \tag{2}$$

where Ω denotes the total failure region. If $x \in \Omega$, $I(x \in \Omega)$ is 1, otherwise it is 0. If there are multiple performances of interest, a failure occurs when at least one of them violates the corresponding specification. Then there will be multiple failure regions, and the total failure region is $\Omega = \bigcup_{k=1}^{N} \Omega_k$ where N is the number of performances of interest.

An SRAM array usually consists of M columns and each of the columns consists of N bit-cells as shown in Fig. 1. To read information saved in the first row, we pre-charge the bit lines set and enable the corresponding word line. If any voltage in the set of differential bit line voltages is smaller than the input offset voltage of sense amplifier in a given delay, the read operation fails.

B. Monte Carlo and Importance Sampling

MC method is the most common and traditional approach to estimate SRAM failure rate. With a mass of samples drawn from the given PDF of variables, the failure rate can be probability estimated as

$$\hat{P}_{MC} = \frac{1}{N} \sum_{i=1}^{N} I(\mathbf{x}_i) \tag{3}$$

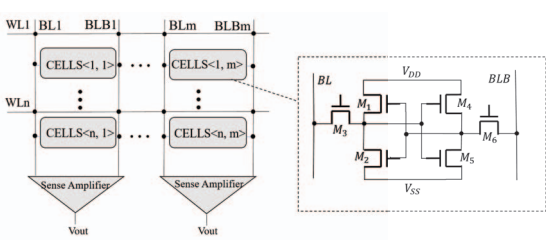

Fig. 1. An SRAM array with sense amplifiers.

where x_i is the i_{th} sample, N is the total sampling amount which is often between 10^7 and 10^8.

Sampling efficiency can be improved by IS method [8]. The main idea of IS-based method is to replace the original distribution $p(x)$ by practical distribution $q(x)$. With N samples generated from the distorted PDF $q(x)$, failure rate P^{fail} can be estimated as

$$P^{fail} \approx \hat{P}_{IS} = \frac{1}{N} \sum_{i=1}^{N} \frac{I(x_i)p(x_i)}{q(x_i)} \tag{4}$$

In yield estimation of post-simulation for large SRAM circuits, there are multiple performances of interest which can induce multiple failure regions. MFRIS [6] can address the yield estimation problem in both high-dimensional and multiple-failure-region cases. MFRIS is an IS-based method whose main contribution is multi-failure-region OSV searching. Searching OSV is a crucial step in most IS-based methods, which is equivalent to solve the following optimization problem

$$\begin{aligned} &min \ \ \|v\| \\ &s.t. \ I(\mathbf{v})=1 \end{aligned} \tag{5}$$

where v is in parameter space, $\| \cdot \|$ denotes the L2-norm. MFRIS extends the idea of MNIS [8]. First, it proposes a Multi-Starting-Point Sequential Quadratic Programming (MSP-OSV) framework to search for OSVs. Next, it constructs a Gaussian mixture distorted distribution based on the multi-region-OSVs. Finally, IS method is used to estimate the total failure rate.

C. Extension of Gaussian Process for variable fidelity data sources

Gaussian Process (GP) regression [9] defines a supervised learning problem, which is an effective method, especially for a moderate size of data sets. We assume that dataset X,Y is generated by an unknown mapping $f(x)$

$$Y = f(X) \tag{6}$$

A typical GP model is defined by a mean function μ and a covariance matrix K:

$$f(x) \sim gp(\mu(X), K(X)) \tag{7}$$

where $X = \{x_1, ..., x_n\}$ represents a training data set of size N, μ is the mean function which is usually set as $\mu = 0$ for simplicity, while K is a kernel matrix.

Some previous works known as Multi-Fidelity methods extend the GP regression framework to construct probabilistic models that enable the combination of variable fidelity data sources. Nonlinear Autoregressive Gaussian Process (NARGP) [7] supposes that there are two fidelity levels of data. X_l and Y_l are the input and the output of the low-fidelity level, while X_h and Y_h are the input and the output of the high-fidelity level. The relationship between the two different fidelity models can be expressed as

$$f_h(x) = z_l(f_l(x)) + \delta_h(x) \tag{8}$$

where $f_h(x)$ and $f_l(x)$ are GP models at the high-fidelity level and the low-fidelity level, z_l represents a mapping from the low-fidelity model to the high-fidelity model. NARGP

can learn nonlinear correlations between various information sources. Because $f_l(x)$ is also a GP model, the function is composited of two GP models. Therefore, we call the function *deep GP*. With an assumption that z_l and δ_h are independent as mentioned in [10], the scheme can be written as

$$f_h(x) = g_h(x, f_l(x)) \tag{9}$$

where g_h represents a GP model over variables vector x and the predict value of the low-fidelity model $f_l(x)$.

III. Proposed Approach

A. Multi-Fidelity Gaussian Process model

Since an SRAM array is a regular matrix-like structure, intuitively, there exist some relationships among performances of different scale SRAM arrays. Fig. 2 gives concrete experimental results to show the correlations of performances on different cases, where the small and the large SRAM arrays are 1x4 and 8x4 respectively. Parameters x1 and x2 are respectively threshold voltage (Vth) of the transistor M2 at $cell < 1, 2 >$ and $cell < 1, 3 >$ as shown in Fig. 1, and the performance is the minimum voltage difference of bit lines among all 4 columns. It can be clearly concluded that performances of different arrays have similar shape over parameters.

Based on such similarity, we introduce MFGP model. The high-fidelity data are from the layouts of required large SRAM arrays. In order to obtain the low-fidelity data, the mutual information (MI) dimensional reduction method as mentioned in [3] is applied for all columns of the large SRAM array first. The MI of two discrete random variables X and Y is defined as

$$I(X,Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right) \tag{10}$$

where p(x,y) is the joint distribution and p(x), p(y) are the marginal distributions. In this work, the same nonparametric method as [3] is applied to calculate MI. Then, bit-cells consisting of the dimensions with the highest MI will be considered to be important. We select rows and columns only including important bit-cells as the low-fidelity data sources.

Since the simulation time of small SRAM arrays is ignorable, we can build GP models on small arrays easily with plenty of samples. Then a nonlinear mapping from the low-fidelity models to the high-fidelity models is built with another GP model. In this way, one can use much fewer samples to build high-fidelity models, which is the basic idea of MFGP. Considering that yield estimation of SRAM with multiple columns on post-simulation is a multiple-failure-region problem, in this work, we construct different MFGP models in different failure regions. Since the process of OSV searching can be very time-consuming, the two fidelity models and OSVs are all undiscovered and need to be modified iteratively.

B. Iterative strategy

It is unnecessary to build accurate MFGP models over all domain of large SRAM arrays. Following the idea of IS-based methods, one only needs to build models accurately near the OSVs. However, The OSVs of large scale circuits are still unknown and calculating them directly is time-consuming. Therefore, an iterative strategy is applied.

First, the OSVs of small arrays should reasonably be good initials of the ones of large arrays. A 2D example of failure boundaries and OSVs of post-simulations is shown in Fig. 3, where the red and blue lines are the failure boundaries of 1x8 and 4x8 SRAM arrays, respectively. OSVL1 and OSVL2 are the two low dimensional OSVs, while OSVH1 and OSVH2 are high dimensional. The directions of low and high dimensional OSVs are very similar, although their lengths are different. Therefore, for the initialization, MFRIS is applied with enough samples and negligible time on small arrays, to calculate low dimensional OSVs and set them as the initials of high dimensional OSVs. Meanwhile, for each low dimensional OSV of multiple failure regions, we sample adequately on small arrays as initial low-fidelity datasets.

For each iteration, for high-fidelity data, a small number of samples around the current OSVs are added to the data set. With the corresponding low-fidelity data set, the MFGP model is updated. Then the new high dimensional OSVs are solved by MSP-SQP optimization solver on the current MFGP model instead of real SPICE simulations. The iteration repeats until the OSVs converge. Because of the prior knowledge from low-fidelity data, the high dimensional OSV will not be far off the real even in early iterations.

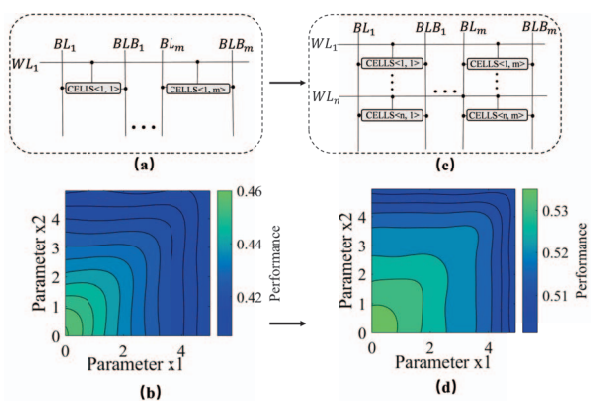

Fig. 2. Colormap for performances and parameters of small and large SRAMs.

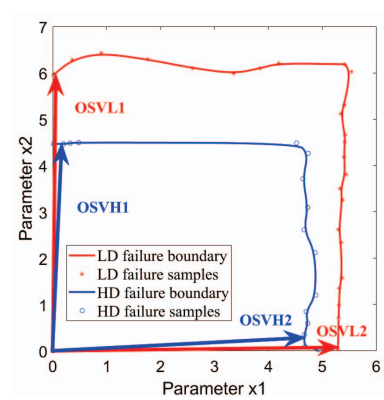

Fig. 3. A case of failure boundaries and OSVs for small and large SRAMs.

C. Prior knowledge of Multi-Modal method

We process a MM method on small scale SRAMs (low dimensional) to further accelerate convergence. The basic idea in this process is to find prior knowledge about the failure boundary and bring the information to the iterative process.

To start the iteration, one needs to sample for initial data set, which is termed *cool start* in GP modeling [11]. Generally, one can simply sample near the current OSVs as shown in Fig. 4(a). Although it works, it is not efficient enough. Intuitively, the samples located near the boundaries are more informative as shown in Fig. 4(b). Therefore, a MM optimal problem [12] is formulated to obtain more informative samples. Then, the failure indication function of the low-fidelity is formulated as

$$\max_{XS \subseteq D} |XS| \tag{11}$$

where $|.|$ is the size of a set, $XS = \{x_i^{lo}, i = 1, 2, ..., n\}$, and x_i^{lo} is the local optimum of the black-box function

$$y = \left| perf^L - spec \right| \tag{12}$$

where $perf^L$ is the performance of the low dimensional SRAM and *spec* is the threshold performance for failure. The local optimal of this black-box function will be failure boundaries of small SRAM circuits. The optimal object of the MM problem is trying to find the local minimal points as many as possible. In this work, a widely-used multi-modal solver NMMSO [13] is applied and the optimal results of MM will be added to the initial data sets which can accelerate the convergence of OSV searching distinctly.

D. Failure rate calculation

After the iteration stops, we can calculate the total failure rate by an efficient IS-based method, i.e., sample around OSVs and calculate the performances from the MFGP model without any extra SPICE simulation. If there are M failure regions with the corresponding OSV s_i^*, we employ Multi-Mean-Shift Importance Sampling (MMSIS) as [6]. Finally, we use (4) to calculate the total failure rate.

$$q(x) = \sum_{i=1}^{M} w_i p(x - s_i^*) \tag{13}$$

$$w_i = \frac{\exp(-\frac{1}{2}||s_i^*||^2)}{\sum_{i=1}^{M} \exp(-\frac{1}{2}||s_i^*||^2)} \tag{14}$$

E. Experimental setting for small scale SRAM arrays

In yield estimations, a SPICE transient simulation is invoked, where the voltage difference y_h of two bit lines of an SRAM column is measured at a given time t_h. If the voltage difference y_h is less than the given specification y_{h0}, it fails. Usually, the measure time t_h is located on the slope of the voltage difference as the black line shown in Fig. 5.

However, in actual post-simulation, the latency of the large SRAM arrays is much longer than the time of the small SRAM arrays, as the red line shown in Fig. 5. If the measure time t_h of the large SRAM arrays is simply applied for the small SRAM arrays, i.e., t_l=t_h, the voltage difference of the small arrays y_l is near to 1, where the features of failure boundary will be a little far away from the slope phase. Therefore, an artificial setting of measure time t_l for the small SRAM arrays can be chosen to make the features of low-fidelity data more similar to high-fidelity data.

The post transient simulations of the small and the large SRAM arrays at Typical-Typical (TT) corner are invoked firstly. For example, Fig. 5 gives the voltage differences of bit lines of the 7th column on 4x8 and 1x8 SRAM circuits. We choose $y_{l0} = y_{h0}$, and t_l can be measured and set in the slope phase as shown in Fig. 5. In this way, failure boundaries of small arrays will be more possibly like the ones of large arrays.

F. The framework of the proposed method

The framework of the proposed method is shown in Fig. 6. First, the dimensional reduction is applied for the large SRAM array with MI method, and the layout of the small SRAM array is drawn. Then, we run post-simulation of the large and the small layouts at TT corner and obtain the measure time t_l and voltage difference y_{l0} of the small array.

On the small SRAM, apply MFRIS for searching low dimensional OSVs and solve MM optimization problem to find enough failure boundary points. On the large SRAM, iterative steps of building MFGP models on performances and searching OSV are executed. Sample near current OSVs, call SPICE to obtain performances, and add these samples into data sets. Update the MFGP model, and solve the updated OSVs with MSP-SQP solver. This iteration will not stop until the convergence of OSV searching. Finally, the yield is estimated based on the OSVs and MFGP models without any realistic SPICE simulations.

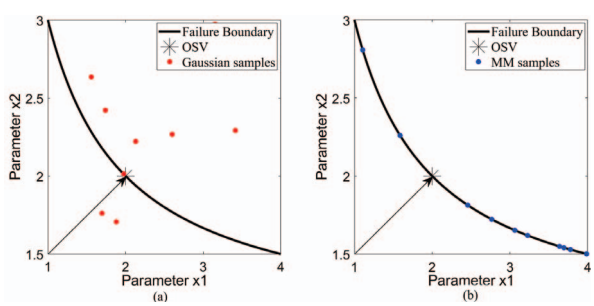

Fig. 4. Illustration of prior knowledge from MM.

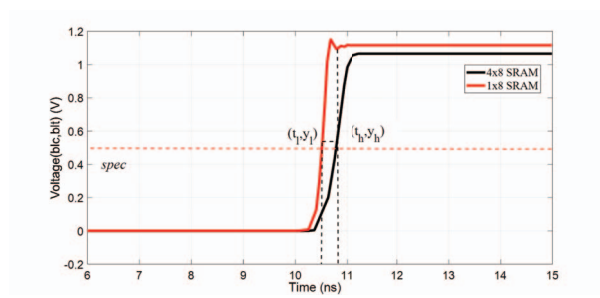

Fig. 5. Measure time setting for the small SRAM circuits.

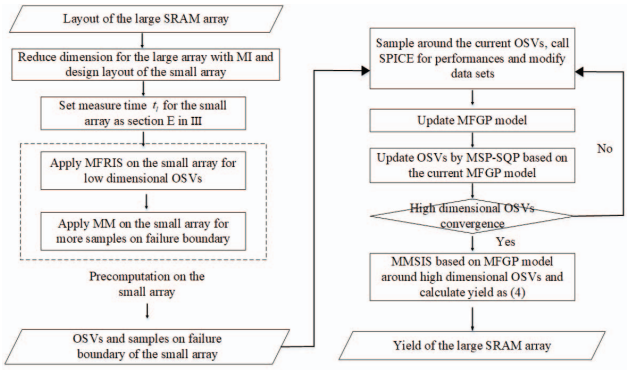

Fig. 6. Framework of the proposed method.

IV. Numerical Experiments

To verify the accuracy and efficiency of the proposed method, the multi-region failure rate estimations of large SRAM arrays on post-simulation are evaluated. All test cases are under a 28nm CMOS process.

MC needs more than 10^7 samples, which is unaffordable for realistic SPICE simulation of large SRAM arrays. Golden results are from MFRIS [6] with enough samples for its simplicity and clarity in theory. Three different approaches, i.e., MFRIS, DAC18 [3], and the proposed method, are implemented in MatLab.

A. Estimation of the read failure rate for post-simulation

Table I summarizes the results of the read operation for an 8x4 SRAM array (8×4×6=192D). DAC18 seems less efficient than the original paper because OSV searching of multiple failure regions dominates the cost. The proposed method needs 340 (100 for dimensional reduction) samples, which is 3.0x faster than the state-of-the-art method in DAC18, with 2.4% relative error. Certainly, the proposed method needs extra simulations on the small 1x4 SRAM array (1×4×6=24D) with 1.3 hours of CPU time, ignorable compared with total cost.

We further verify the proposed method with an 8x8 SRAM array (8×8×6=384D) as shown in Table II. In this case, the proposed method can be 5.3x faster than DAC18, with 2.7% relative error. The improvement on effciency is much more important than the improvement on accuracy which makes the proposed method acceptable for industrial application.

B. Estimation of the write failure rate for post-simulation

Table III summarizes the results of the write operation for an 8x8 SRAM array. Note that in SRAM arrays, the write operation mode is very different from the read mode, while our proposed method is still efficient. The proposed method can be 6.9x faster than DAC18, with 3.7% relative error.

TABLE I
Accuracy and Speed Comparisons for Read Failure Rate of an 8x4 SRAM Array on Post-Simulation

	Failure Rate (E-05)	Relative Error (%)	#Samples of 192D	Total CPU Time (h/1c)	Speed Up
Golden	5.31	N/A	7560	630.0	1.0
MFRIS	5.87	10.5	1626	135.5	4.6
DAC18	4.57	13.9	1026	85.0	7.3
Proposed	5.18	**2.4**	**100 + 240**	28.0	22.2

TABLE II
Accuracy and Speed Comparisons for Read Failure Rate of an 8x8 SRAM Array on Post-Simulation

	Failure Rate (E-05)	Relative Error (%)	#Samples of 384D	Total CPU Time (h/10c)	Speed Up
Golden	2.98	N/A	4400	366.7	1.0
MFRIS	2.65	11.0	2398	199.8	1.8
DAC18	2.63	11.7	1798	150.2	2.4
Proposed	2.90	**2.7**	**340**	28.4	12.9

C. Estimation of the failure rate for pre-simulation

The proposed method can also work on pre-simulation cases. Experimental results with a 32x1 SRAM array on pre-simulation are shown in Table IV. In this case, there is only one performance of interest, leading to one failure region. Therefore, the superiority of our proposed method is not as distinct as the multiple failure region cases. However, the proposed work can still gain 5.0x speed up over MFRIS and 1.7x speed up over DAC18.

D. Computational complexity vs circuit scale

The computational complexity is evaluated with the growth of both rows and columns. As mentioned in [3], only a few variables have an impact on performance, which makes the number of important variables grows very slightly with rows. Therefore, the computational complexity of both proposed method and DAC18 can be roughly constant to the increase of rows as shown in Fig. 7(a). we start from an SRAM array with 4 rows and 1 column only (24D) and add the number of rows to 64 (384D).

However, with the grow of the column number, there will be more failure regions as well as influential variables. This increases simulations for both MSP searching process and SQP process in OSV searching, which will domain the total cost

TABLE III
Accuracy and Speed Comparisons for Write Failure Rate of an 8x8 SRAM Array on Post-Simulation

	Failure Rate (E-06)	Relative Error (%)	#Samples of 384D	Total CPU Time (h/10c)	Speed Up
Golden	5.74	N/A	3833	319.8	1.0
MFRIS	6.07	5.7	2746	229.1	1.4
DAC18	5.27	8.1	2206	184.0	1.7
Proposed	5.53	**3.7**	**320**	26.7	12.0

TABLE IV
Accuracy and Speed Comparisons for Read Failure of a 32x1 SRAM Column on Pre-Simulation

	Failure Rate (E-05)	Relative Error (%)	#Samples of 192D	Total CPU Time (h/1c)	Speed Up
Golden	3.33	N/A	2368	197.3	1.0
MFRIS	3.22	3.3	1069	89.0	2.2
DAC18	3.13	6.0	369	30.7	6.4
Proposed	3.45	**3.6**	**210**	17.5	11.3

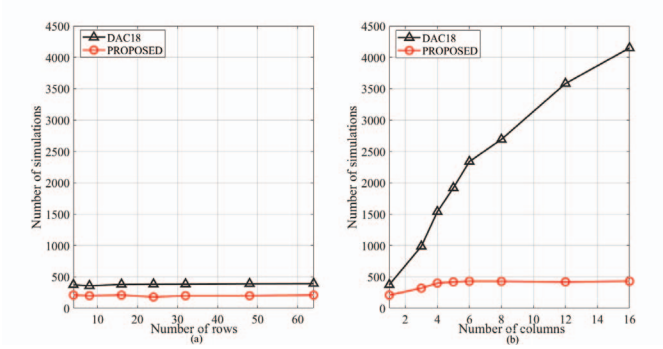

Fig. 7. Number of simulations versus rows and columns.

on large scale cases. Therefore, the total cost will be roughly linear to the increase of columns for DAC18. As shown in Fig. 7(b), we start from an SRAM array with 4 rows and 1 column only (24D) and increase the number of columns to 16 (384D). With the prior knowledge and the iterative strategy, simulations for OSV searching are needless. Therefore, the cost of the proposed method can grow very slightly with columns, which is very appealing for large arrays.

E. Robustness of OSV searching

We further verify the robustness of iterative OSV searching for large arrays. Fig. 8(a) is the standard convergence process where the initial point is the OSV of the small array. In Fig. 8(b), the initial OSV is given about 25% disturbance. As shown in Fig. 8, the proposed iterative OSV searching strategy can converge rapidly and accurately in both cases.

F. Effect of Multi-Modal method

The optimal results of MM method contain more information of failure boundaries, which makes them more valuable than the naive Gaussian samples. The experimental results based on a 4x8 SRAM array are shown in Table V, while the golden total failure rate is simulated by MFRIS with sufficient samples. Adding 30 extra samples, Gaussian samples have no improvement on the relative error, while MM method can improve 29% (38.1%-9.1%) of the relative error.

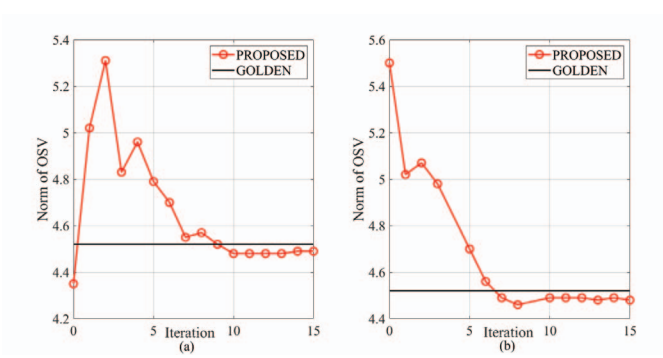

Fig. 8. Robustness of OSV searching in large SRAM arrays.

TABLE V
EFFECT OF MULTI-MODAL METHOD FOR FAILURE RATE ESTIMATION

	Failure Rate (E-06)	Relative Error	Total Samples
Golden	7.44	N/A	8755
No extra samples	4.60	38.1%	300+0
Gaussian samples	4.06	45.4%	300+30
MM samples	6.76	**9.1%**	300+30

V. CONCLUSIONS

In this paper, we firstly propose an efficient method to handle yield estimation problem on post-simulation for SRAM with multiple columns. We take advantage of low dimensional data adequately in the process of both searching OSV and building MFGP models. Experimental results verify that the proposed method can earn 5-7x speedup with higher accuracy over the state-of-the-art method with 384D cases.

ACKNOWLEDGEMENTS

This work was supported in part by the National Key Research and Development Program of China under Grant 2020YFA0711901, in part by National Natural Science Foundation of China (NSFC) research projects 61974032, 61774045, 61929102, 61674042.

REFERENCES

[1] Solido Design Automation Inc., "High-sigma monte carlo for high yield and performance memory design," Tech. Rep., 2011.
[2] S. Sun and X. Li, "Fast statistical analysis of rare circuit failure events via subset simulation in high-dimensional variation space," in *ICCAD*, 2014.
[3] J. Zhai, "An efficient bayesian yield estimation method for high dimensional and high sigma sram circuits," in *DAC*, 2018.
[4] L. Dolecek, M. Qazi, D. Shah, and A. Chandrakasan, "Breaking the simulation barrier: Sram evaluation through norm minimization," in *ICCAD*, 2008.
[5] J. Yao, Z. Ye, and Y. Wang, "Importance boundary sampling for sram yield analysis with multiple failure regions," *TCAD*, vol. 33, no. 3, pp. 384–396, 2014.
[6] M. Wang, C. Yan, X. Li, D. Zhou, and X. Zeng, "High-dimensional and multiple-failure-region importance sampling for sram yield analysis," *TVLSI*, vol. 25, no. 3, pp. 806–819, 2017.
[7] P. Perdikaris, M. Raissi, A. Damianou, N. D. Lawrence, and G. E. Karniadakis, "Nonlinear information fusion algorithms for data-efficient multi-fidelity modelling," *Proc Math Phys Eng Sci*, 2017.
[8] 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 *DATE*, 2010.
[9] C. E. Rasmussen, *Gaussian Processes in Machine Learning*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 63–71.
[10] L. Le Gratiet, "Recursive co-kriging model for design of computer experiments with multiple levels of fidelity," *Int J Uncertain Quantif*, vol. 4, 10 2012.
[11] K. Swersky, J. Snoek, and R. P. Adams, "Multi-task bayesian optimization," in *NIPS*, 2013.
[12] J. E. Fieldsend, "Running up those hills: Multi-modal search with the niching migratory multi-swarm optimiser," in *CEC*, 2014.
[13] G. Singh and K. Deb, "Comparison of multi-modal optimization algorithms based on evolutionary algorithms," in *GECCO*, 2006.
[14] J. Tao, H. Yu, D. Zhou, Y. Su, X. Zeng, and X. Li, "Correlated rare failure analysis via asymptotic probability evaluation," in *DAC*, 2017.
[15] X. Jing and R. Yao, "A fast-simulation model for post-layout sram," in *ICASIC*, 2007.
[16] M. Kennedy and A. O'Hagan, "Predicting the output from a complex computer code when fast approximations are available," *Biometrika*, vol. 87, no. 1, pp. 1–13, 2000.
[17] B. C. Ross, "Mutual information between discrete and continuous data sets," *PLOS ONE*, vol. 9, no. 2, pp. 1–5, 02 2014.