Efficient Yield Analysis for SRAM-Based System with PDF Consolidation Methodology

Shih-Han Chang, Ling-Yen Song, Yen-Chen Chun, Yu-Cheng Tsai and Chien-Nan Liu

Institute of Electronics, National Yang Ming Chiao Tung University, Hsinchu City, Taiwan, R.O.C.
Email: shchang.ee09@nycu.edu.tw; audrey1535.ee07g@nctu.edu.tw; erica66119@gmail.com; rambo890316@gmail.com; jimmyliu@nycu.edu.tw

Abstract - SRAM-based system is one of the most popular design in various applications. However, the simulation cost for yield estimation is often very high due to the high yield requirement of SRAM circuits. Importance sampling techniques are able to reduce the number of samples in high sigma analysis. However, the complexity is still high if the entire memory system with peripheral circuits are simulated together. To handle this issue, we propose an efficient yield analysis method for the overall SRAM system. Instead of analyzing the whole system directly, the proposed methodology evaluates each circuit block first. Then, the interactions of circuit blocks are considered to evaluate the system performance accurately with the prior distribution of each block. In this way, the overall accurate yield estimation can be obtained easily. The experimental results demonstrate that the proposed methodology efficiently estimates the yield of SRAM-based designs with high accuracy, especially for rare events.

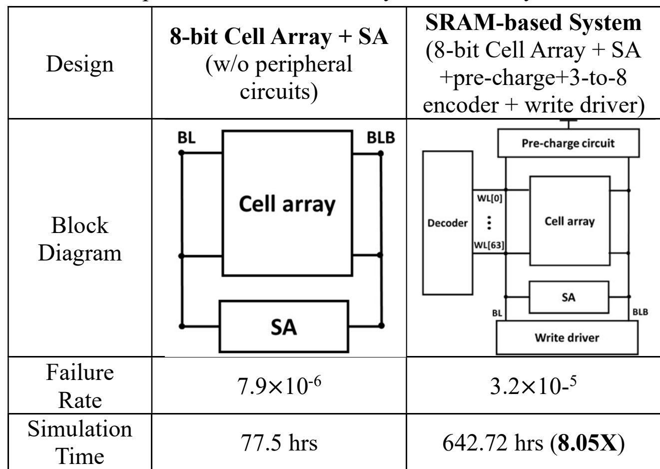
I. Introduction

As the increasing demand for Internet of Things (IoT) and machine learning applications, SRAM circuit becomes one of the critical blocks in the system IC design [1]. Generally, SRAM macro is usually well-designed by foundry with high yield performance. However, designers usually add or change the peripheral circuits in SRAM system to achieve the specific targets in different applications [1]. Even the yield of SRAM macros is high enough, the SRAM-based system with new peripheral circuits usually cannot remain the high quality, as shown in Fig. 1. Therefore, to keep sufficient reliability for the SRAM-based system, the overall design yield still requires to be analyzed and optimized in sizing loops.

A straightforward approach to observe the impact of process variation on Performance of Interest (PoI) is adopting multiple sampling simulations, such as Monte-Carlo (MC) analysis [2–3]. The circuit yield can be estimated by the output probability density function (PDF). However, because SRAM-based system usually consists of millions of transistors, MC method becomes inefficient for such high dimensional analysis. Moreover, the yield of SRAM system is so high that the rare failure events can be detected by running roughly 10^8 random samples at least. Therefore, as shown in TABLE I, MC analysis requires very high simulation cost, which is impractical for the products with time-to-market pressure.

Since the conventional MC method is computationally expensive for searching the rare events, importance sampling (IS) [4] techniques have been proposed to reduce the computational cost. The key idea of the importance sampling method is to reduce the required number of samples by shifting the sampling distribution toward the boundary of device parameters, which increases the probability of hitting the failure region. By using the importance sampling, more samples are drawn at the tail part of the yield distribution. However, in IS method, the complexity of determining the likely failure region is still high for large circuits due to the exploration of the complicated parameter space, which can only be obtained by analyzing bit-cell or cell array. In [5], a qualitative statistical analysis technique has been proposed to estimate the yield of SRAM macros. To consider the correlation between SA and cell array, the cumulative distribution function of each cell is assumed as independent and identical, which may cause the large yield accuracy loss. Therefore, an efficient and accurate yield analysis method is required to deal with high dimensional and high-sigma properties of SRAM-based systems.

TABLE I
Comparison between cell array and overall system

Design	**8-bit Cell Array + SA** (w/o peripheral circuits)	**SRAM-based System** (8-bit Cell Array + SA +pre-charge+3-to-8 encoder + write driver)
Block Diagram		
Failure Rate	7.9×10^{-6}	3.2×10^{-5}
Simulation Time	77.5 hrs	642.72 hrs (**8.05X**)

In this paper, we propose an efficient yield analysis approach for SRAM-based systems by the distribution consolidation process. With the pre-evaluated yield performance of each block, we consider the correlation between each block and calibrate the failure rate of each block according to the output signals of preceding block to obtain precise system design yield. Since the yield of each block can be obtained with lower simulation cost, the overall simulation cost of system yield analysis can be greatly reduced by the proposed approach. Moreover, because of the prior individual yield evaluation, it is easy to analyzed which block is the

bottleneck of system yield that needs further optimization. If there is any modification, the system yield is able to be obtained by just recompiling the distribution consolidation process rather than re-evaluating all of blocks in the system. In this way, redundant simulations can be greatly reduced for the refined system, which improves the efficiency of yield analysis in the system design flow. While compared to the MC analysis, the proposed approach can reduce 83.29% and 88.84% runtime for read and write stability analysis respectively in the experiments, with little sacrifice on yield accuracy.

The rest of this paper is organized as follows. Section II briefly introduce some preliminaries of yield analysis and the operations of SRAM-based system. The proposed system yield estimation methodology is presented in Section III. The experimental results on the operations of SRAM-based system are provided in Section IV to demonstrate the accuracy and efficiency of the proposed approach. Finally, some conclusions are drawn in Section V.

II. Process Variation in SRAM-based System

A. Failure Rate Evaluation

To analyze the design yield, the variations of device parameters should be considered in the circuit simulation. Without loss of generality, process parameters are supposed to be mutually independent, which is usually modeled by Gaussian distribution in (1). μ is the mean value of the distribution, and σ is the standard deviation of the distribution.

$$\mathcal{N}(\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{v-\mu}{\sigma}\right)^2\right) \quad (1)$$

The conventional method to estimate the yield is using MC method. It is the most straightforward and widely used approach, and is often considered as the golden answer due to its high accuracy. To estimate the PDFs of PoIs $Y_{MC}(x)$ with the random variable x, we have to generate N random samples for simulation in the MC method. After simulating all samples, the indicator function is introduced to represent whether this sample falls into the failure region or the acceptable region. Assume the threshold for the performance metric is Y_{thr}, then the indicator function can be defined as (2). Then, the yield is the probability of the samples falling into the acceptable region, which can be expressed as (3). And the probability of failure rate can be obtained by (4).

$$\mathrm{I}(x_i) = \begin{cases} 1, & y_{MC}(x_i) > y_{thr} \\ 0, & y_{MC}(x_i) < y_{thr} \end{cases} \quad (2)$$

$$Y_{MC}(x) = \frac{1}{N}\sum_{i=1}^{N} I(x_i) \quad (3)$$

$$P_{f_{MC}} = 1 - Y_{MC}(x) \quad (4)$$

The accuracy of the probability density function depends on the number of samples. For a high-sigma analysis, most samples will fall into the acceptable region. Thus, if the number of samples is not enough, none of the samples will fall into the failure region. However, it doesn't mean that the yield is 100%. To ensure the estimated distribution with (1-ε) accuracy and (1-δ) confidence, the required number of samples N(ε,δ) can be determined by the function shown in (5).

$$\mathcal{N}(\varepsilon, \delta) \approx \frac{\log\left(\frac{1}{\delta}\right)}{\varepsilon^2 P_{f_{MC}}} \quad (5)$$

For example, if both the accuracy and confidence level are required to achieve 95%, and the failure probability is about 10^{-6} for a circuit design with high robustness (e.g., SRAM circuit), the required number of simulation samples in the Monte Carlo method should be more than 10^8. It is impractical to simulate so many samples, especially for a large-scaled circuit. Moreover, if the peripheral circuits around the memory array are also considered, the requirements for simulation resource will become even larger. Apparently, using Monte Carlo method to analyze the yield is infeasible due to the resource limitation.

B. Interactive Effects Analysis in SRAM-based System

Process variations may impact the characteristics of transistors, thus inducing the variation of circuit performance. PoI is the performance used to judge the functional correctness or whether the specification is met. In the yield analysis of system design, the PoIs are the signals that communicate between sub-circuits. According to the interactions of sub-circuits in the read operation, whose signal direction is shown in Fig. 1(a), the PoI of pre-charge circuit is the bit-line voltage that can be charged. It determines the initial values of the bit-lines in the cell array. However, the pre-charge circuit is not considered in most existing works, in which the initial bit-lines voltages are assumed to be ideal at the supply voltage. PoI of the decoder is the pulse duration time for turning on the pre-charge circuit or word-line, which determines the required time to discharge one of the bit-lines. Offset voltage is the PoI of SA, which determines the correctness of functionality.

Take the write operation of a SRAM circuit as an example, the overall signal path is shown in Fig. 1(b). The failure of write operation occurs if the data is not flipped after the word line pulse duration. Since the circuit behaviors are the same in read and write operations, PoI of the pre-charge circuit and the decoder are also same. Because process variation may lead to the different discharge ability of a write driver, PoI of the write driver is the voltage of discharged bit-line.

III. SRAM-based Yield Estimation Methodology

The proposed efficient yield analysis flow for SRAM-based system is shown in Fig. 2. The typical SRAM-based system is partitioned as five functional blocks: write buffer, decoder, pre-charge circuit, SA and cell array. After obtaining the PDF of each block, the distribution consolidation process is applied to predict the yield of two blocks with signal interaction between them. The consolidation operations are orderly applied by following the signal path of SRAM-based system until the read/write operation is completed. The final failure rate is obtained after the last block is finished. The details of this process are explained as follows.

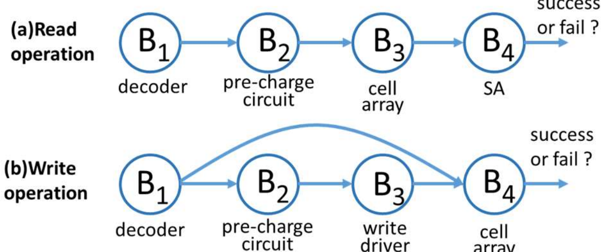


Fig. 1. Signal paths of the (a)read and (b)write operations

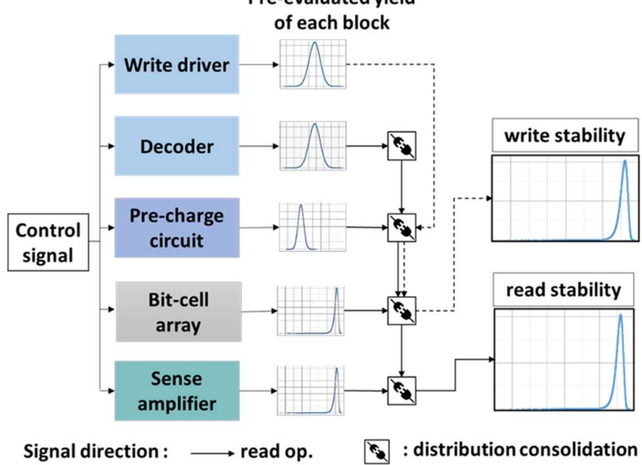


Fig. 2. The proposed fast yield analysis for SRAM-based system

A. Block-level Yield Analysis

When we design the system circuit, each circuit block is often designed first instead of designing the whole system circuit directly. When we analyze each sub-circuit separately, the interactive effects between sub-circuits are not considered. Without the interactive effects, all sub-circuits are assumed to be independent of each other. While considering interactive effects, the simple Gaussian distribution can be used as the input PDFs for the block-level yield analysis. However, the real distribution may not be the same as the assumed PDF, which will cause the predictive yield loss. Therefore, after the output PDFs of each sub-circuit are obtained, they will be used as the new input PDFs of the next stage sub-circuit. The original output PDF is consolidated to produce an accurate distribution by using the proposed transfer function. The modified output PDF will be the input distribution of the next stage. The consolidation process will be performed iteratively until obtaining the primary output yield of the system.

B. Transfer Function for Distribution Consolidation

The overall distribution consolidation process is shown in Fig. 3. In the block-level yield analysis, the raw output PDFs are produced, which are the initial distributions have not considered interactive effects yet. That is, the circuit property is already recorded by sufficient samples. Next, to consolidate the two raw output PDFs, a fine-tune layer is added after the individual circuit analysis to reshape the PDF of the succeeding stage while considering the interactive effects between two circuit blocks. In order to enhance the efficiency

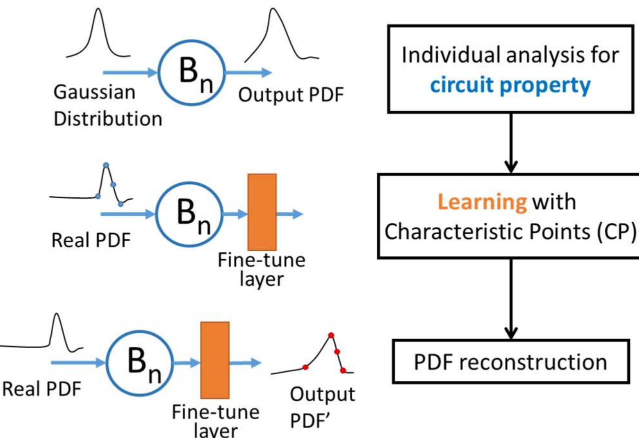


Fig. 3. Overall flow of distribution consolidation.

of training the fine-tune layer, we concentrate the samples of a specific region into a super point, which is called as characteristic points (CPs) in this work. Via allocating sampling resources on CPs, we can ensure sufficient data on the critical tail region of PDF while training the fine-tune layer. Lastly, by calibrating every sample in the raw PDF with the trained fine-tune layer, the raw PDF is able to be reconstructed into a new PDF that considers the interactive effects between two circuit blocks.

In the block-level yield analysis, we use a Gaussian distribution as the input noise of the block B_1, which is shown in the orange PDF curve $f_{in}(B_1)$ in Fig. 4, to obtain the raw output PDF $f_{out}(B_1)$ in Fig. 4. The orange dots in Fig. 4 depict the data samples generated in block-level analysis. As the previous block B_2 has finished the block-level analysis, the $f_{in}(B_1)$ will be updated by $f_{in}(B_{12}) = f_{out}(B_2)$, and the output should be altered to $f_{out}(B_{12})$ as well. Instead of simulating the whole system, we use the ML model to predict the output movement of samples directly. Our ML model has two layers, the circuit performance evaluating layer and the input noise fine-tune layer. In the first layer, the preliminary output performance will be obtained based on the knowledge learned in block-level analysis. In the second fine-tune layer, the impact coming from the noise of previous stage will be augmented to the output result. The fine-tune layer consists of two parts, the arrangement and learning kernels. In the arrangement, data will be allocated to different predictive kernels according to their output level of CP. We use the first-order regression model as the predict kernel in this paper to determine the magnitude of output movement.

After the consolidation step, the sub-system PDF of B_1 and B_2, $f_{out}(B_{12})$, can be evaluated easily by inference with the changed input $f_{out}(B_1)$ the raw PDF $f_{out}(B_2)$. In inference process, each previous sample *s* is adjusted by two information, the difference between input noise and the performance change rate estimated by fine-tuning layer, which are depicted as input and output movements in Fig. 4. $f_{out}(B_{12})$ can be formed, and the failure rate can be determined easily with these updated samples.

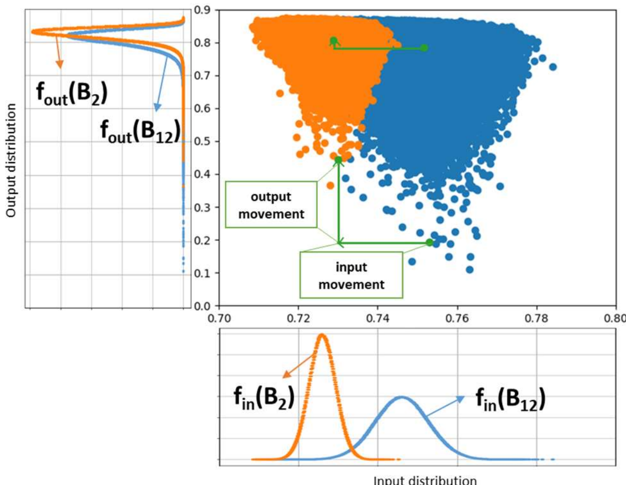


Fig. 4. PDF adaptation considering interactive effect.

IV. Experimental Result

In the experiments, the 64-bit 6T-SRAM memory array with peripheral circuits including encoder/decoder, write buffer and SA is used to demonstrate the proposed design yield estimation method. The variation of threshold voltage in each transistor is 20%, which is set by a gauss function with 6 sigma analyses in HSPICE simulator. The accuracy of the proposed approach is verified by Monte Carlo analysis on the whole SRAM design directly, which is regarded as Golden in the following. Because of rare failure event, 107 simulations are required in each Monte Carlo analysis with 95% accuracy and 95% confidence level and extra 30 samples for training the scaling factor. All the experiments are performed by using Synopsys HSPICE and executed on an Intel Xeon Gold 6248 CPU at 2.5GHz with 186GB memory.

Table II shows the required time and the obtained accuracy of each yield analysis approach for read/write stability. Compared to the golden result, the mathematical approach [5] reduces simulation time for read/write stabilities by 54%, but the predicted yield is 10 times higher. In our approach, the required time can be reduced by 88%. When compared to [5], the required time can also be reduced by 75%. During read operations, the proposed approach is able to provide very similar yield result compared to the golden result. Fig. 5 shows the three output distributions produced by the three approaches. The result demonstrates that our method can still predict accurate system yield on these rare failure events. While reducing the computational cost, the estimation accuracy can still be kept in our approach.

TABLE II
The comparison between golden, [3] and this work.

	Read Stability			Write Stability		
methodology	Golden	[5]	**This work**	Golden	[5]	**This work**
Failure rate	2.97×10^{-4}	3.99×10^{-3}	**2.90×10^{-4}**	2.27×10^{-4}	3.44×10^{-3}	**2.01×10^{-4}**
Sim. Time (hr.)	2364	1084	**395**	2401	1087	**268**
Time of math. model	-	< 1s	**< 10s**	-	< 1s	**< 10s**
Speedup	1	54.15%↓	**83.29%↓**	1	54.72%↓	**88.84%↓**
	-	1	**63.56%↓**	-	1	**75.34%↓**

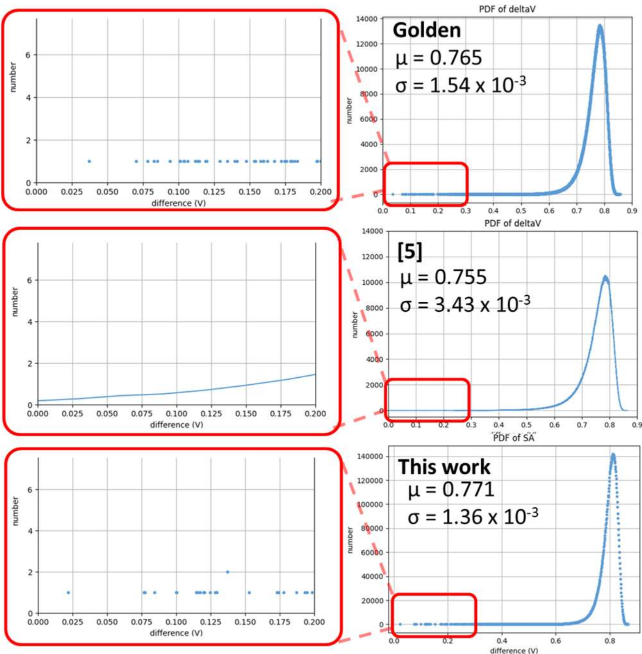


Fig. 5. The comparisons of output distributions for read stability.

V. Conclusions

In this paper, we propose an efficient yield analysis methodology to consider the impact of peripheral circuits in SRAM-based system design. After analyzing each block in the system separately, we consolidate the output distribution stage-by-stage using the proposed distribution consolidation approach. The transfer learning technique is used to adapt the PDF with the proposed characteristic points to deal with the non-Gaussian tail accurately. Then, the overall system yield can be estimated without expensive simulations of the entire system design. Furthermore, after redesigning or substituting the sub-circuits, the system yield of modified system design can be easily obtained by the proposed consolidation process since only the modified blocks need to be re-analyzed. Therefore, the proposed methodology significantly improves the efficiency of yield estimation in the SRAM-based system design flow.

References

[1] W. Choi and J. Park, "An efficient convolutional neural networks design with heterogeneous SRAM cell sizing," International SoC Design Conference, 2017.

[2] K. D. Dingeç and W. Hörmann, "Improved Monte Carlo and quasi-Monte Carlo methods for the price and the Greeks of Asian options," Winter Simulation Conference, 2014.

[3] I. Han, L. Yu and Y. Shin, "Fast Monte Carlo method via reduced sample number and node filtering," International Conference on Integrated Circuit Design and Technology, 2010.

[4] M. Rakka and R. Kanj, "Importance Splitting Sample Point Reuse for Efficient Memory Yield Estimation, "International Symposium on Circuits and Systems, 2021.

[5] P. Sharma, A. K. Gundu and M. S. Hashmi, "Modeling and yield estimation of SRAM sub-system for different capacities subjected to parametric variations," Int. Symp. VLSI Design and Test, 2016.