A Fast and Provably Bounded Failure Analysis of Memory Circuits in High Dimensions

Wei Wu†, Fang Gong†, Gengsheng Chen*, Lei He†

†Electrical Engineering Dept., UCLA
Los Angeles, CA 90095, US
{weiwu2011,gongfang}@ucla.edu
lhe@ee.ucla.edu

*School of Microelectronics, Fudan Univ.
Shanghai, 201203, China
gschen@fudan.edu.cn

Abstract— Memory circuits have become important components in today's IC designs which demands extremely high integration density and reliability under process variations. The most challenging task is how to accurately estimate the extremely small failure probability of memory circuits where the circuit failure is a "rare event". Classic importance sampling has been widely recognized to be inaccurate and unreliable in high dimensions. To address this issue, we propose a fast statistical analysis to estimate the probability of rare events in high dimensions and prove that the estimation is always bounded. This methodology has been successfully applied to the failure analysis of memory circuits with hundreds of variables, which was considered to be very intractable before. To the best of our knowledge, this is the first work that successfully solves high dimensional "rare event" problems without using expensive Monte Carlo and classic importance sampling methods. Experiments on a 54-dimensional SRAM cell circuit show that the proposed approach achieves 1150x speedup over Monte Carlo without compromising any accuracy. It also outperforms the classification based method (e.g., Statistical Blockade) by 204x and existing importance sampling method (e.g., Spherical Sampling) by 5x. On another 117-dimension circuit, the proposed approach yields 364x speedup over Monte Carlo while existing importance sampling methods completely fail to provide reasonable accuracy.

I. INTRODUCTION

Memory circuits (e.g., SRAM bit-cell, sense amplifier, delay chain, etc.) need to be replicated millions or even billions of times to achieve extremely high integration density in a smaller footprint, where the cutting-edge process technology is demanded. In this case, the stringent yield requirement of memory circuits can be translated into an extremely small failure probability of each component circuit, thereby making the circuit failure a "rare-event" [1].

In general, the probability estimation of "rare-event" is usually analytical intractable due to high complexity of memory circuits, therefore, sampling methods must be used. The most straightforward approach is the Monte Carlo (MC) method, which repeatedly draws samples and evaluates circuit performance with transistor-level SPICE simulation. However, MC is extremely time-consuming for rare-event estimation, because millions or even billions of samples are needed to capture one single failure.

To mitigate the complexity issue of the MC method, many statistical methodologies have been developed in past few years [2, 3, 4, 5, 6, 7, 8] which can be categorized into two groups:
(1) **Classification**: the approach in [2] makes use of a "classifier" to "block" those Monte Carlo samples that are unlikely to cause failures and simulates the remaining samples. However, this method has two limitations: first, a perfectly accurate classifier is usually unavailable. A safety margin is used in [2] to prevent the classifier error. Second, the imperfect classifier can easily incur large error beyond the safety margin for circuits with irregular failure region and strongly nonlinear behavior, which typically cannot be detected by the approach in [2].
(2) **Importance Sampling**: several approaches in [3, 4, 5, 6, 7] had been developed to construct a new "proposed" sampling distribution under which a "rare event" becomes "less rare" so that more failures can be easily captured. The critical issue is how to build an optimal proposed sampling distribution. Previous work investigated different approaches. For example, [3] mixes a uniform distribution, the original sampling distribution and a "shifted" distribution centering around the failure region. The approaches in [4, 5] simply shift the sampling distribution towards the point of failure region with a minimum L_2-norm. The work in [6] uses "particle filtering" to tilt more samples towards the failure region. The approach in [7] approximates the optimal sampling distribution with a parameterized sampling distribution by minimizing the Kullback-Leibler (KL) distance between them. These importance sampling based methods are plagued by the curse of high dimensionality [9, 10, 11]. In general, they can only be used in low-dimensional problems (e.g., those with a scope of 6-12 variables) but become very untrustworthy for high-dimensional problems.
Clearly, most of existing approaches can successfully be applied to low-dimensional problems with a few random variables but, in general, perform poorly in high dimensions. Therefore, an effective and low-complexity approach is still urgently needed for failure analysis of memory circuits in high dimensions.

In this paper, we proposed a novel statistical algorithm to efficiently estimate the failure probability of memory circuits in high dimension, where tens or hundreds of random variables are present. In details, the proposed methodology first constructs a new subset of the sampling space that dominates the failure region for memory circuits and can be efficiently estimated with a few samples. Then, the failure probability of memory circuits can be evaluated by the product rule of conditional probability within this sampling subset space. More importantly, the estimation from the proposed method is proved to be always bounded in high dimensions. Experiments on a 54-dimensional SRAM cell circuit show that the proposed approach achieves 1150X speedup over Monte Carlo without compromising any accuracy. It is also 204X faster than the classification based method (e.g., Statistical Blockade [2]) by and 5X faster than existing importance sampling method (e.g., Spherical Sampling [4, 5]). On another 117-dimension circuit, the classification based method fails to block "unlikely to fail" samples, and Spherical Sampling [4, 5] method completely fails to provide reasonable accuracy. Contrastingly, the proposed approach yields accurate result with 364X speedup over Monte Carlo.

The rest of this paper is organized as follows. In Section 2, we provide the necessary background on importance sampling and revisit the reasons for its failure in high dimensions. Section 3 describes the techniques underpinning the proposed algorithm in detail. The experiments are provided in Section 4 to validate the accuracy and efficiency of proposed method. This paper is concluded in Section 5.

978-1-4799-2816-3/14/$31.00 ©2014 IEEE

II. Background

A. Formulation of Probability Estimation

Let $f(X)$ be a probability density function (PDF) for a random variable X (e.g., any process or electronic variable parameters) which is the input of a measurement process as shown in (1); the output Y is an observation (e.g., voltage, amplitude, period, etc.) with input X:

$$\underbrace{X}_{\text{variable}} \Rightarrow \boxed{\text{Measurement, SPICE, etc.}} \Rightarrow \underbrace{Y}_{\text{observation}} \quad (1)$$

Usually, it is of great interest to estimate the probability of Y from a small subset $\mathcal{S}$ of the entire sampling space. For example, a small subset is the "failure region" for SRAM design and includes all failed samples where performance constraints cannot be satisfied. Therefore, the probability $p(Y \in \mathcal{S})$ can be estimated as:

$$p(Y \in \mathcal{S}) = \int I(X) \cdot f(X) dX. \quad (2)$$

$$I(X) = \begin{cases} 0 & \text{if } Y \notin \mathcal{S} \\ 1 & \text{if } Y \in \mathcal{S} \end{cases}$$

where Y is the observation/performance with the input variable X and the indicator function $I(\cdot)$ identifies whether $Y \in \mathcal{S}$ or not. Note that the integral in equation (3) is intractable because the analytical formula of $I(X)$ is unavailable. Therefore, sampling based method must be used. For example, the MC method enumerates as many samples of X as possible (e.g., $x_1, \cdots, x_n$) according to $f(X)$ and evaluates their indicator function values to estimate $p(Y \in \mathcal{S})$ as:

$$\tilde{p}(Y \in \mathcal{S}) = \frac{1}{n}\sum_{i=1}^{n} I(x_i) \xrightarrow[n\to+\infty]{a.s.} p(Y \in \mathcal{S}). \quad (3)$$

Here $\tilde{p}(X \in \mathcal{S})$ is an unbiased estimate from sampling method and can be very close to $p(X \in \mathcal{S})$ with a large number of samples.

B. Importance Sampling (IS)

When $Y \in \mathcal{S}$ is a *rare event*, the MC method becomes extremely inefficient because most $I(x_i)$ are zeros. Millions or billions of samples of X are needed to capture only one failed sample from the failure region $\mathcal{S}$.

To deal with this issue, the *importance sampling* (IS) has been introduced to sample from a "proposed" sampling distribution $g(X)$ that tilts towards $\mathcal{S}$ where a rare-event becomes more likely to happen:

$$\begin{aligned} p_{IS}(Y \in \mathcal{S}) &= \int I(X) \cdot \frac{f(X)}{g(X)} \cdot g(X) dX \\ &= \int I(X) \cdot w(X) \cdot g(X) dX. \end{aligned} \quad (4)$$

Here, $w(X)$ is the "likelihood ratio" or the weight for each sample of X. $w(X)$ compensates for the discrepancy between $f(X)$ and $g(X)$ and unbiases the probability estimation under $g(X)$. Sampling based methods can be used to evaluate above integral as:

$$\tilde{p}_{IS}(Y \in \mathcal{S}) = \frac{1}{n}\sum_{j=1}^{n} w(\tilde{x}_j) \cdot I(\tilde{x}_j) \xrightarrow[n\to+\infty]{a.s.} p(Y \in \mathcal{S}). \quad (5)$$

$\tilde{x}_j$ $(j = 1, \cdots, n)$ follows the "proposed" sampling distribution $g(X)$ rather than the original distribution $f(X)$, because more *rare event* samples in the subset $\mathcal{S}$ can be easily chosen under the distribution $g(X)$.

Theoretically, $\tilde{p}_{IS}(Y \in \mathcal{S})$ is consistent with $p(Y \in \mathcal{S})$ in (3) if $\text{supp}(g(X)) \supset \text{supp}(I(X) \cdot f(X))$, where $\text{supp}(\cdot)$ denotes the support of a probabilistic distribution.

C. Failure Analysis of Importance Sampling

While importance sampling is, in principle, mathematically correct, the *degeneration* or *collapse* of the likelihood ratios leads to the failure of importance sampling in high dimensions as discussed in [10, 11].

Let us consider a classical case, as shown in Fig. 1, where $f(X)$ is the "original" sampling distribution and $g(X)$ is the "proposed" sampling distribution. The small circles with the same size within $g(X)$ are samples drawn from $g(X)$. In the bottom of Fig. 1, a few circles with different sizes represent the illustrative scales of the likelihood ratios corresponding to the samples on top of them. Clearly, if $g(X)$ has thinner tails than $f(X)$, the likelihood ratios $w(X) = f(X)/g(X)$ approach infinity in the tails of $g(X)$. Hence, the likelihood ratios vary dramatically and have extremely large variance that leads to unstable probability estimate.

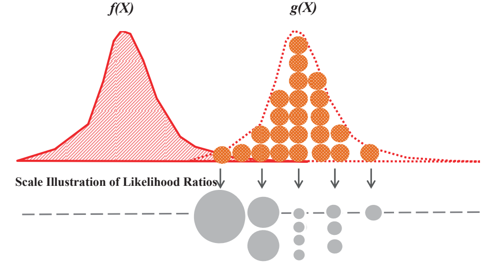


Fig. 1. The scale illustration of likelihood ratios in importance sampling.

Moreover, the reason for the collapse of likelihood ratio can be explained from another perspective: when importance sampling shifts $g(X)$ towards the rare-event region that is typically in the tails of $f(X)$, $f(X)$ and $g(X)$ become mutually singular and have "disjoint" support [10]. Therefore, IS fails to retain its accuracy.

This collapse issue of likelihood ratios becomes much worse in high dimensions because $w(X)$ is a product of probabilities for multiple parameters and consequently approaches infinity more quickly.

III. Proposed Method

A. Algorithm Overview

We consider a small subset $\mathcal{S}$ as the failure region in SRAM design under the given performance constraint (e.g., the performance of SRAM circuit Y should be greater than certain performance threshold t_c). Hence the subset $\mathcal{S} = \{Y|Y \geq t_c\}$ in Fig. 2 contains all failed samples that are "rare events".

The basic idea of the proposed algorithm is to construct a new subset $\mathcal{T}$ with a new threshold t (e.g., $t = 0.99$-quantile point). This new subset $\mathcal{T} = \{Y|Y \geq t\}$ includes "non-rare" events and dominates the "rare event" subset $\mathcal{S}$ (e.g., $\text{supp}(\mathcal{T}) \supset \text{supp}(\mathcal{S})$).

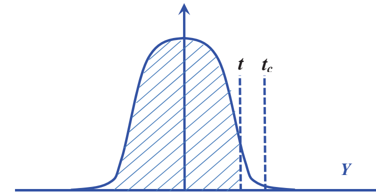


Fig. 2. Basic idea in proposed algorithm. (Noted that $\mathcal{T} = \{Y|Y \geq t\}$ contains $\mathcal{S} = \{Y|Y \geq t_c\}$).

In this way, the failure probability of SRAM design can be estimated by a product rule from the probability theory [12]:

$$P(Y \geq t_c) = P(Y \geq t) \cdot P(Y \geq t_c | Y \geq t). \quad (6)$$

The proposed algorithm has two stages and can be illustrated with Fig. 3:

1) Initial Sampling with MC: This step aims to evaluate the probability $P(Y \in \mathcal{T}) = P(Y \geq t)$ where t is the threshold, such as $t = 0.99$-quantile point shown in the left of Fig. 3. Since the samples in $\mathcal{T}$ are "non-rare" events, this evaluation needs only a few samples using standard MC method.

2) Conditional Probability Estimation: The most challenging task is to efficiently evaluate the conditional probability $P(Y \geq t_c | Y \geq t)$ where sampling method must be used. To expedite the convergence rate of estimation, a "proposed" sampling distribution $g(X)$ that is close to the failure region shall be constructed by *shifting* and *reshaping* the

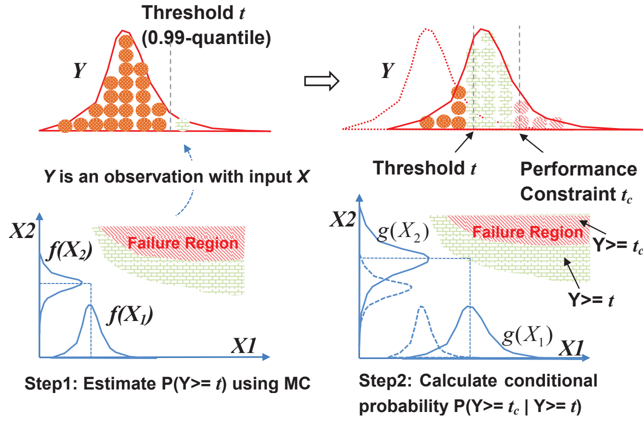


Fig. 3. Overall flow in proposed algorithm. (Noted that $\mathcal{T} = \{Y|Y \geq t\}$ contains $\mathcal{S} = \{Y|Y \geq t_c\}$).

Algorithm 1 Overall Algorithm

Input: random variables X with sampling distributions $f(X)$ and performance constraints $Y \geq t_c$.
Output: the estimation of failure probability $p_{IS}(Y \geq t_c)$.

1: **/* 1: Initial Sampling with MC */**
2: Use few MC samples to find the threshold value t of performance (e.g., t = 0.99-quantile point).
3: Run standard Monte Carlo method to calculate $P_{MC}(Y \geq t)$ with certain accuracy level.
4: **/* 2: Conditional Probability Calculation */**
5: Shift the original sampling distribution $f(X)$ towards the failure region.
6: Reshape the shifted $f(X)$ by changing its standard deviation to construct $g(X)$.
7: Generate samples from $g(X)$ and evaluate conditional probability $P(Y \geq t_c|Y \geq t)$.
8: **/* 3: Failure Probability Estimation */**
9: Solve for the failure probability $p_{IS}(Y \geq t)$ as

$$P_{IS}(Y \geq t_c) = P_{MC}(Y \geq t) \cdot P(Y \geq t_c|Y \geq t).$$

"original" sampling distribution (shown in the right of Fig. 3). More details will be discussed in the following section.

The overall algorithm flow is described in Algorithm(1). There are several issues that need to be resolved: 1) It is, at the moment, unclear how to *shift* and *reshape* the original sampling distribution $f(X)$ in order to build $g(X)$; 2) With the proposed sampling distribution $g(X)$, how to calculate the conditional probability; 3) It is of great interest to study whether the estimations of proposed algorithm is always bounded or not.

The following sections discuss how we solve these issues.

B. Shift and Reshape Sampling Distribution

B.1 Mean-Shift Vector Selection

Mean-shift is a typical way to move the sampling distribution towards the failure region where the failed samples are most likely to happen in previous works [3, 4, 5, 6, 7]. The key is to find the mean-shift vector for the original sampling distributions $f(X)$.

To this end, we propose to shift $f(X)$ towards a "non-rare" subset $\mathcal{T} = \{Y|Y \geq t\}$, because our target is to evaluate the conditional probability $P(Y \geq t_c|Y \geq t)$ around the subset $\mathcal{T}$. More importantly, as $\mathcal{T}$ is usually not far away from the mean of $f(X)$, the shifted distribution shares almost the *same* support with $f(X)$ so as to avoid the "disjoint support" issue.

In addition, we adopt the insights from [7] to find a close-to-optimal mean-shift vector in this work. Let us consider a 1-D problem as an example. The algorithm in [7] starts with an initial parameterized distribution $\hat{f}(X, \hat{\mu})$ and tries to update the mean value iteratively to achieve a close-to-optimal sampling distribution $f^*(X, \mu^*)$ by an analytic formula:

$$\mu^* = \frac{\sum_{i=1}^{N} I(x_i) \cdot w(x_i) \cdot x_i}{\sum_{i=1}^{N} I(x_i) \cdot w(x_i)}. \tag{7}$$

Here x_i $(i = 1, \cdots, N)$ are samples drawn from $\hat{f}(X, \hat{\mu})$ and $w(x_i)$ are their likelihood ratios as $w(x_i) = f(x_i)/\hat{f}(x_i, \hat{\mu})$.

Intuitively, the updated mean value μ^* can be viewed as the coordinates of the *centroid point* in the failure region where the failed samples are most likely to happen. This interesting finding becomes more obvious if $\hat{f}(X, \hat{\mu})$ equals $f(X)$ and all likelihood ratios take on value 1. Hence, μ^* is:

$$\mu^* = \frac{\sum_{i=1}^{N} I(x_i) \cdot x_i}{\sum_{i=1}^{N} I(x_i)}. \tag{8}$$

Therefore, our mean-shift method tries to shift the sampling distribution towards the "centroid point" of the subset $\mathcal{T} = \{Y|Y \geq t\}$, which can be evaluated with available MC samples from the first step in Algorithm (1) and requires no extra sampling/simulation cost.

B.2 Standard Deviation Selection

Next, it is desired to *reshape* the shifted sampling distribution around the centroid of subset $\mathcal{T}$. In particular, the standard deviation for the proposed sampling distribution $g(X)$ must be properly chosen to reach the failure region $\mathcal{S} = \{Y|Y \geq t_c\}$, because the shifted and reshaped sampling distribution should *dominate* or completely cover the "rare-event" region $\mathcal{S}$.

As an illustration, let us consider a 2-D problem in Fig. 4. The problem now becomes how to choose the standard deviation of the proposed sampling distribution $g(X)$ to obtain the samples in the "rare-event" region $\mathcal{S} = \{Y|Y \geq t_c\}$.

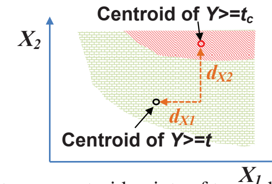


Fig. 4. The distance between centroid points of two subsets along each parameter axis.

The proposed algorithm first approximates the centroid point of $\mathcal{S} = \{Y|Y \geq t_c\}$ using uniformly-distributed samples and then calculates the distance between these two centroid points along each parameter axis (e.g., d_{X_1} and d_{X_2} shown in Fig.4). Then, we choose $max(d_{X_i}, \sigma_{(0,X_i)})$ as the standard deviation of $g(X_i)$ for the variable X_i, where $\sigma_{(0,X_i)}$ is the original standard deviation of $f(X_i)$. This choice can be intuitively explained as follows:

- $d_{X_i} > \sigma_{(0,X_i)}$**:** the failure region $\mathcal{S}$ is very far away from the subset $\mathcal{T}$, therefore, the larger value d_{X_i} is used to extend the range of $g(X_i)$ and obtain the rare-event samples in the failure region. In the meantime, $g(X_i)$ has almost the same supports with $f(X_i)$ because its mean position locates at the centroid point of $\mathcal{T}$ and is not far away from $f(X_i)$.
- $d_{X_i} < \sigma_{(0,X_i)}$**:** Suppose the smaller one, d_{X_i}, is chosen as the standard deviation of $g(X_i)$, the proposed sampling distribution $g(X)$ will have much smaller sampling space, thereby, making it fail to keep the same supports with $f(X_i)$ and suffer from "disjoint supports" issue. The proposed algorithm chooses $\sigma_{(0,X_i)}$ as the standard deviation of $g(X_i)$ in this case.

C. Conditional Probability Calcualtion

With the proposed sampling distribution $g(X)$, it is desired to efficiently estimate the conditional probability in Algorithm(1). We can start with the product rule in the probability theory [12]:

$$P(Y \geq t_c|Y \geq t) = \frac{P(Y \geqslant t_c, Y \geqslant t)}{P(Y \geqslant t)}. \tag{9}$$

In addition, when samples x_i $(i = 1, \cdots, N)$ are generated from $g(X)$, both $P(Y \geqslant t_c)$ and $P(Y \geqslant t)$ can be estimated mathematically with the indictor function and likelihood ratios. Thus, the equation (9) becomes:

$$P_{MIS}(Y \geq t_c | Y \geq t) = \frac{P(Y \geqslant t_c)}{P(Y \geqslant t)} = \frac{\frac{1}{N}\sum_{i=1}^{N} w(x_i) \cdot I_{\{Y \geqslant t_c\}}(x_i)}{\frac{1}{N}\sum_{i=1}^{N} w(x_i) \cdot I_{\{Y \geqslant t\}}(x_i)}. \quad (10)$$

where $I_{\{Y \geqslant t_c\}}(\cdot)$ and $I_{\{Y \geqslant t\}}(\cdot)$ are indicator functions for subsets $Y \geq t_c$ and $Y \geq t$, respectively. $w(x_i)$ are likelihood ratios for these samples. In this way, the conditional probability can be efficiently evaluated under proposed sampling distribution $g(X)$.

D. Boundedness Analysis

D.1 Importance Sampling

Let us first investigate the existing importance sampling and assume samples x_j $(j = 1, \cdots, M)$ are generated from the proposed sampling distribution $g(X)$.

We find the upper bound of probability estimate from the conventional importance sampling according to Boole's inequality (also known as the union bound from probability theory [12]) as:

$$P(Y \geqslant t_c) = P_f(\sum_{j=1}^{M} I_{\{Y \geqslant t_c\}}(x_j)) \leqslant \sum_{j=1}^{M} P_f(x_j) \cdot I_{\{Y \geqslant t_c\}}(x_j) = \sum_{j=1}^{M} w(x_j) \cdot I_{\{Y \geqslant t_c\}}(x_j). \quad (11)$$

In 11 P_f stands for the probability estimation under sampling distribution $f(X)$. As discussed in [10, 11], the likelihood ratios $w(x_j)$ can vary dramatically in high dimension and be any random quantities. Therefore, the union bound of the estimation $P(Y \geqslant t_c)$ in (11) approaches infinity and importance sampling becomes unreliable and untrustworthy.

D.2 Proposed Algorithm

The proposed algorithm constructs a subset $\mathcal{T} = \{Y | Y \geq t\}$ that *dominates* the failure region $\mathcal{S} = \{Y | Y \geq t_c\}$ (i.e., $\mathcal{T} \supset \mathcal{S}$). Therefore, the upper bound of conditional probability can be derived as:

$$P(Y \geqslant t_c | Y \geqslant t) = \frac{P(Y \geqslant t_c)}{P(Y \geqslant t)} = \frac{\sum_{j=1}^{N} w(x_j) \cdot I_{\{Y \geqslant t_c\}}(x_j)}{\sum_{j=1}^{N} w(x_j) \cdot I_{\{Y \geqslant t\}}(x_j)} \leqslant 1. \quad (12)$$

Note that no matter how likelihood ratios $w(x_j)$ vary, the same likelihood ratios for samples in the failure region $\mathcal{S} = \{Y | Y \geq t_c\}$ would appear in both numerator and denominator in (12) if and only if the calculations of both $P(Y \geqslant t_c)$ and $P(Y \geqslant t)$ utilize the *same* set of samples x_j $(j = 1, \cdots, M)$ drawn from $g(X)$. Clearly, the conditional probability estimation of proposed algorithm is always bounded by the upper bound 1. Thereby, the propose algorithm can reliably provide bounded estimation results.

IV. Experimental Results

We investigate its performance of the proposed algorithm for failure analysis of memory circuits (e.g., SRAM bit-cell and sense amplifier) in this section. All experiments are performed using MATLAB and Hspice with BSIM4 transistor model. The proposed algorithm is named as HDIS (high-dimensional importance sampling) in this section. In addition, Monte Carlo (MC), statistical blockade (SB)[2], and spherical sampling (SS) [4, 5] have been implemented for comparison purpose.

TABLE I
PROCESS PARAMETERS OF MOSFETS.

Variable Name	σ/μ	unit
Flat-band Voltage (V_{fb})	0.1	V
Gate Oxide Thickness (t_{ox})	0.05	m
Mobility (μ_0)	0.1	m^2/Vs
Doping concentration at depletion (N_{dep})	0.1	cm^{-3}
Channel-length offset (ΔL)	0.05	m
Channel-width offset (ΔW)	0.05	m
Source/drain sheet resistance (R_{sh})	0.1	Ohm/mm^2
Source-gate overlap unit capacitance (C_{gso})	0.1	F/m
Drain-gate overlap unit capacitance (C_{gdo})	0.1	F/m

A. SRAM Circuit and Variation Modeling

A functional diagram of SRAM circuit with one bit-cell column is shown in Fig. 5, which consists of a decoder, bit-cells, a sense amplifier and a delay chain [13]. During the reading operation: the bit-cells store the data in forms of '0' or '1'; the decoder generates an address of a specific bit-cell and releases a read enable signal. Therefore, the chosen bit-cell starts to discharge the bit-lines (i.e., the lines that connect to all bit-cells) to produce a voltage difference between two bit-lines. the sense amplifier reads out the stored data by capturing and magnifying the voltage difference on bit-lines.

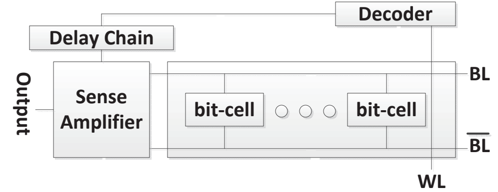


Fig. 5. Functional diagram of an SRAM circuit.

The process variations are introduced into each transistor of SRAM circuit, which are modeled by 9 process parameters shown in Table (I). The parameters are physically independent [14] and can be considered to be Gaussian random variables. Note that the threshold voltage V_{th} is not a process parameter and depends on V_{fb}, t_{ox}, ΔL and ΔW through related effects [14].

B. SRAM Cell with Reading Failure

A typical 6-transistor SRAM bit-cell is shown in Fig. 6: $Mn2$ and $Mn4$ control the accessing of the cell; the remaining four transistors form two inverters and use two stable states (either '0' or '1') to store the data in this memory cell. The *reading access failure* happens when the voltage difference between $\bar{BL}$ and BL is too small to be sensed by the sense amplifier at the end of reading operation [1].

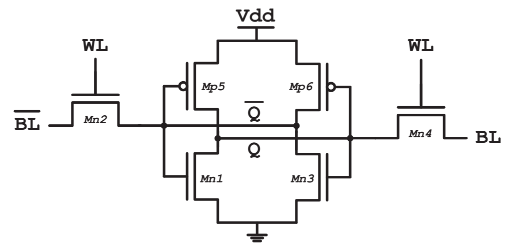


Fig. 6. The schematic of the 6T SRAM cell.

We perform different methods (MC, SS[5], SB[2], proposed) on this SRAM bit-cell example to predict the reading failure probability under process variations and the comparison results are shown in Table II.

B.1 Accuracy Comparison

At a first glance, we would be very surprised to find that SS[5] method based on conventional importance sampling framework can provide accurate failure rate predictions in this 54-dim problem!

TABLE II
COMPARISON FOR SRAM BIT-CELL ANALYSIS WITH 90% TARGET ACCURACY AND CONFIDENCE LEVEL.

	Monte Carlo (MC)	Spherical Sampling (SS)[5]	Statistical Blockade (SB)[2]	Proposed method (HDIS)
failure probability	2.413E-05 (0%)	2.8415E-05 (+17.7%)	2.7248e-05 (+12.9%)	2.4949E-05 (+3.39%)
#sim. runs	4.6e+6 (1150X)	2e+4 (5X)	8.16e+5 (204X)	4e+3 (1X)

However, this comparison cannot allow us to reach that conclusion, because this SRAM bit-cell example is a "pseudo" high-dimensional problem for two-fold reasons: (1) during the reading operation, not all transistors are active. In fact, both $Mp5$ and $Mn3$ are shut off, therefore, the process variations on these two transistors have no effect on discharge behavior of bit-lines at all; (2) without loss of generality, assuming $\bar{BL}$ = '0' and BL='1', the discharge current flows from $\bar{BL}$ to the ground through $Mn2$ and $Mn1$ so that to pull down the voltage of $\bar{BL}$. As such, the process variations in $Mn2$ and $Mn1$ have more significant effects on the discharge behavior of bit-lines and can potentially mask the variation effects in $Mp6$ and $Mn4$. In this way, there are only 18 "effective" variable parameters, which suggests that this example is a problem with modest dimension.

When compared with MC results, the proposed method provides the most accurate failure probability estimation with only 3.39% relative error, while the estimations from SS[5] and SB[2] have more than 10% relative error.

B.2 Efficiency Comparison

From Table II we also compare the efficiency of these methods: MC is very time-consuming and requires nearly 4.6 millions transistor-level SPICE simulations; SB[2] can provide 6X complexity reduction by screening out and simulating those "most-likely-to-fail"samples; SS[5] method is made more efficient (230X speedup over MC) by better choosing failed samples using importance sampling algorithm; the proposed algorithm achieves the best convergence rate (1150X faster than MC) by efficiently spreading more samples into the failure region using a sampling distribution with a large-standard-deviation in high dimensions.

C. Sense Amplifier for Target Gain

Next, we consider a sense amplifier example which includes 13 transistors as shown in Fig. 7.

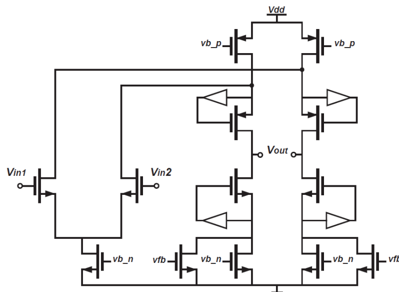


Fig. 7. The schematic of a sense amplifier circuit.

In a SRAM circuit, the sense amplifier is designed to magnify the voltage difference between $\bar{BL}$ and BL. If the gain is too small, the output of this amplifier might be too weak to be read by the decoder circuit. Therefore, a reading failure happens. With the variation modeling summarized in Table I, the sense amplifier example has 117 random variables in total. More importantly, all of these variable parameters are "effective" because the transistors are active and process variations on each transistor can significantly change the gain, which is a truly high-dimensional problem.

C.1 Accuracy Comparison

To validate the accuracy of the proposed algorithm, we apply different methods (MC, SS[5], SB[2] and proposed) on this 117-dim problem to predict the timing failure probability. Here, MC serves as the "gold standard". SB[2], is not included in the further comparison, because the classifier used in SB[2] fail to block any Monte Carlo sample. Therefore, Considering the complexity of running the classifier, the SB [2] involves even higher computation complexity than MC method.

The evolution of the probability estimation in different methods are plotted in Fig. 8(a). Several observations can be made:

First, this figure shows the failure of conventional importance sampling (i.e., SS[5]). In fact, due to the degeneration or collapse of likelihood ratios, SS[4, 5] method converges to a random quantity which is obviously wrong and far away from the MC result. Moreover, SS[5] does not have a mechanism for improving accuracy even though more samples are added.

The proposed method builds an effective proposed sampling distribution to choose more failed samples easily and its estimation is theoretically bounded due to the proposed evaluation of conditional probability. Therefore, the proposed algorithm can reliably estimate the failure probability that matches with MC results.

C.2 Efficiency Comparison

Even though the Fig. 8(a) provides a rough comparison of efficiency, the detailed comparison can be shown in Fig. 8(b), where different methods try to achieve the "comparable" accuracy. Note that circuit simulation is the most time-consuming part and the runtime cost of the remaining computation becomes negligible. As such, the required number of circuit simulations for the same accuracy and confidence level serves as a measurement of the efficiency.

First, the Figure-Of-Merit (FOM) is used to quantify the accuracy of probability estimation as [4, 5]:

$$\rho = \frac{\sqrt{\sigma^2_{p(\mathrm{fail})}}}{p(\mathrm{fail})}. \tag{13}$$

where $p(\mathrm{fail})$ is the failure probability and $\sigma_{p(\mathrm{fail})}$ is the standard deviation of $p(\mathrm{fail})$. In fact, the FOM can be viewed as a *relative error* so that lower FOM means higher accuracy of probability estimation.

We compare the evolutions of FOM for different methods in Fig. 8(b) and draw a dash line to indicate the 90% accuracy with 90% confidence ($\rho = 0.1$). And we can have following observations:

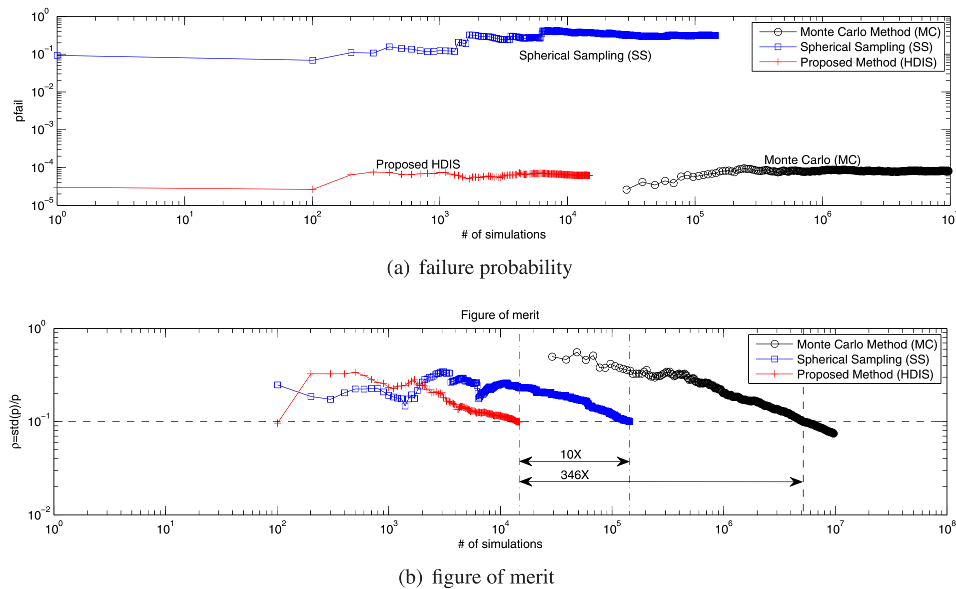
First, SS[5] has reached $\rho = 0.1$ but its estimation is completely wrong. Clearly, it cannot detect the failure at all. The same observation is applied to other existing importance sampling methods due to the boundedness analysis in Section 3.3.1.

Second, The proposed algorithm can provide the accurate estimation of failure probability with only a few thousands samples, which dramatically relieves the requirements of computing and storage efforts. As shown in this figure, the proposed method can achieve $708X$ speedup over Monte Carlo and be $17X$ faster than statistical blockade method [2].

C.3 Comparison for Different Failure Probabilities

We study various methods on the sense amplifier example with three different failure probabilities summarized in Table III. It is obvious that SS[5] method fails to achieve any reasonable accuracy in all these cases. This demonstrates the failure of conventional importance sampling method. On the contrary, the estimates from the proposed method match the MC result.

In addition, the table reveals that the proposed method provides the fastest convergence speed in all these cases and, more importantly, offers substantial complexity reduction as the failure probability becomes



(a) failure probability

(b) figure of merit

Fig. 8. Evolution comparison of the failure probability estimation and figure of merit for different methods.

TABLE III

COMPARISON FOR SENSE AMPLIFIER ANALYSIS WITH 90% TARGET ACCURACY AND CONFIDENCE LEVEL

Target Failure Probability		Monte Carlo (MC)	Spherical Sampling (SS) [5]	Proposed Method (HDIS)
8e-3	prob:(failure)	8.136e-4	0.2603	7.861e-3 (3.4%)
	#sim. runs	4.800e+4 (24X)	16000 (8X)	2000
8e-4	prob:(failure)	8.044e-4	0.2541	8.787e-4 (9.2%)
	#sim. runs	4.750e+5 (36X)	8.330e4 (6.4X)	1.300e4
8e-5	prob:(failure)	8.089e-5	0.3103	8.186e-5 (1.2%)
	#sim. runs	5.156e+6 (346X)	1.430e+5 (10X)	1.500e+4

smaller. This property makes our proposed algorithm suitable for industrial problems where exist “rare events” with extremely small probability.

V. Conclusion

In this paper, we propose a fast statistical algorithm to estimate the extremely small probability of rare events in high dimensions which has proved to be bounded. The proposed algorithm has been successfully applied to failure probability prediction of memory circuits (e.g., SRAM bit-cell, sense amplifier) and demonstrates significant complexity reduction without compromising the accuracy. To the best of our knowledge, this is the first work that successfully handle the rare event estimation in high dimensions without using MC and classic importance sampling method. Experiments on a 54-dimensional SRAM cell circuit show that the proposed approach achieves 1150X speedup over Monte Carlo without compromising any accuracy. It is also 204X faster than the classification based method (e.g., Statistical Blockade [2]) by and 5X faster than existing importance sampling method (e.g., Spherical Sampling [4, 5]). On another 117-dimension circuit, the classification based method fails to improve the performance by blocking “unlikely to fail” samples, and Spherical Sampling [4, 5] method completely fails to provide reasonable accuracy. Contrastingly, the proposed approach yields accurate result with 364X speedup over Monte Carlo.

References

[1] K. Agarwal and S. Nassif, “Statistical analysis of SRAM cell stability,” ser. DAC ’06, 2006, pp. 57–62.

[2] A. Singhee and R. A. Rutenbar, “Statistical blockade: very fast statistical simulation and modeling of rare circuit events and its application to memory design,” *IEEE Tran. on CAD*, vol. 28, pp. 1176–1189, 2009.

[3] 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 *Proceedings of the 43rd annual Design Automation Conference*, ser. DAC’06, 2006, pp. 69–72.

[4] L. Dolecek, M. Qazi, D. Shah, and A. Chandrakasan, “Breaking the simulation barrier: SRAM evaluation through norm minimization,” in *Proceedings of the 2008 IEEE/ACM International Conference on Computer-Aided Design*, ser. ICCAD ’08, 2008, pp. 322–329.

[5] M. Qazi, M. Tikekar, L. Dolecek, D. Shah, and A. Chandrakasan, “Loop flattening and spherical sampling: Highly efficient model reduction techniques for SRAM yield analysis,” in *Design, Automation Test in Europe Conference Exhibition (DATE), 2010*, 2010.

[6] K. Katayama, S. Hagiwara, H. Tsutsui, H. Ochi, and T. Sato, “Sequential importance sampling for low-probability and high-dimensional SRAM yield analysis,” in *IEEE/ACM International Conference on Computer-Aided Design*, 2010.

[7] F. Gong, S. Basir-Kazeruni, L. Dolecek, and L. He, “A fast estimation of SRAM failure rate using probability collectives,” in *Proc. ACM ISPD*, 2012, pp. 41–47.

[8] C. Dong and X. Li, “Efficient SRAM failure rate prediction via Gibbs sampling,” in *Proceedings of the 43rd annual Design Automation Conference*, ser. DAC’11, 2011.

[9] S. Au and J. Beck, “Important sampling in high dimensions,” *Structural Safety*, vol. 25, no. 2, pp. 139 – 163, 2003.

[10] T. B. B. Li and P. Bickel, “Curse-of-dimensionality revisited: Collapse of importance sampling in very high-dimensional systems,” *Technical Report No.696,Department of Statistics, UC-Berkeley*, 2005.

[11] 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, pp. 547 – 568, 2009.

[12] A. Papoulis and S. Pillai, “Probability, random variables and stochastic processes,” *McGraw-Hill*, 2001.

[13] A. Pavlov and M. Sachdev, “CMOS SRAM circuit design and parametric test in nano-scaled technologies: Process-aware SRAM design and test,” *Springer Publisher*, 2008.

[14] P. Drennan and C. McAndrew, “Understanding MOSFET mismatch for analog design,” *IEEE J. of Solid State Circuits*, vol. 38, no. 3, pp. 450 – 456, 2003.