IEICE TRANS. ELECTRON., VOL.E97–C, NO.4 APRIL 2014

PAPER Special Section on Solid-State Circuit Design — Architecture, Circuit, Device and Design Methodology

Hypersphere Sampling for Accelerating High-Dimension and Low-Failure Probability Circuit-Yield Analysis

Shiho HAGIWARA[†a)], **Takanori DATE**[†], *Nonmembers*, **Kazuya MASU**[†], *and* **Takashi SATO**[††b)], *Members*

SUMMARY This paper proposes a novel and an efficient method termed hypersphere sampling to estimate the circuit yield of low-failure probability with a large number of variable sources. Importance sampling using a mean-shift Gaussian mixture distribution as an alternative distribution is used for yield estimation. Further, the proposed method is used to determine the shift locations of the Gaussian distributions. This method involves the bisection of cones whose bases are part of the hyperspheres, in order to locate probabilistically important regions of failure; the determination of these regions accelerates the convergence speed of importance sampling. Clustering of the failure samples determines the required number of Gaussian distributions. Successful static random access memory (SRAM) yield estimations of 6- to 24-dimensional problems are presented. The number of Monte Carlo trials has been reduced by 2–5 orders of magnitude as compared to conventional Monte Carlo simulation methods.

key words: *design for manufacturing, Monte Carlo method, importance sampling, SRAM, process variation, yield, norm minimization, Gaussian mixture models, clustering, hypersphere sampling*

1. Introduction

Technology scaling has brought dramatic improvements in the performance of large scale integrated circuits. Scaling, on the other hand, has introduced process-parameter variations of transistors and interconnections. The influence of process-parameter variations has become one of the most serious concerns in modern circuit designs [1], and it is expected to become even more serious. Random dopant fluctuation is a representative variability [2], which leads to a large random variation in the threshold voltages of transistors. Variation in the threshold voltage is unavoidable in scaled technologies because it is almost impossible to control both the location and the number of dopants. Yield as well as performance optimization under parameter variability are always the principal design objectives.

It is widely known that the yield of a static random access memory (SRAM) is particularly sensitive to the variation in the threshold voltage, because transistors in the SRAM cells are designed using minimum feature sizes [3]. In order to maximize the memory density while maintaining high yield, it is critical to optimize the SRAM cells to achieve very low failure probability. Simulation tool support that facilitates efficient failure rate estimation is necessary.

The Monte Carlo (MC) method [4] is widely adopted for failure probability estimations. An advantage of the MC method is that its accuracy can be improved by increasing the number of MC trials. However, a drawback of this method is that when it is applied to SRAM cells, it becomes time consuming to obtain a reliable estimation. Because SRAM cells have very low failure probability, only a very small fraction of the MC samples falls in the failure region. A new method, which is efficient, accurate, and suitable for analyzing such low failure probability circuits, is required.

There are several methods that attempt to accelerate the SRAM yield analysis [5]–[9]. The authors in [5] proposed to apply the extreme value theory [10]. This theory is effective in dealing with rare probabilities of a continuous value, such as the ones in write-time analysis of an SRAM cell. In [6]–[9], methods based on importance sampling (IS) are used. IS is one of the methods that overcomes the trade-off between the simulation time and accuracy. In the case of IS, samples following an alternative distribution instead of the original distribution are generated. This method increases the number of failure samples, thereby accelerating convergence of the estimation. An appropriate weight is multiplied to each sample to compensate for the estimation bias associated with altering the distributions. More recently, application of sophisticated MC methods is proposed. Examples of such methods include sequential IS [11] and the Markov chain MC method [12].

Among other methods, IS is a simple yet effective method to calculate the expectation of an indicator function. The indicator function indicates whether the given input sample is a pass sample or a failure sample. In particular, the indicator function returns 0 for pass samples and 1 for failure samples. The failure rate of the SRAM cells can be estimated by counting the number of 1's returned by the indicator function. In [9], a method termed norm minimization is proposed, which is based on the large deviations theory [13]. This method reduces the estimation variance by mean-shift IS, in which the mean of the original sample distribution is shifted to the point of a failure sample that has a minimum norm. However, it is difficult to find the minimum norm sample. Prior knowledge of a circuit structure has been used to limit the search space. In addition, the quality of the minimum norm sample may be insufficient, particularly when a small number of samples are used to reduce the search time, which may lead to unstable estimation. A more general, but still an efficient, method that can

Manuscript received July 23, 2013.
Manuscript revised November 26, 2013.

[†]The authors are with the Solutions Research Laboratory, Tokyo Institute of Technology, Yokohama-shi, 226-8503 Japan.
[††]The author is with the School of Informatics, Kyoto University, Kyoto-shi, 606-8501 Japan.
a) E-mail: paper@lsi.pi.titech.ac.jp
b) E-mail: takashi@i.kyoto-u.ac.jp

DOI: 10.1587/transele.E97.C.280

be applied to a variety of problems is required.

In this paper, we propose a novel method termed hypersphere sampling to resolve the above mentioned issues. The proposed method serves as a preprocess for the mean-shift IS that uses Gaussian mixture models as an alternative probability distribution. The appropriate shift vectors are searched efficiently without any special knowledge of the circuit under analysis. The proposed method concentrates samples in the most critical regions, wherein the effect of the parameter that dominates the estimation is observed.

The proposed method involves three steps. In the first step, we extensively search for the failure regions by incremental hypersphere sampling (IHS). In the IHS, MC samples are generated on a hypersphere surface. The radius of the hypersphere is incrementally increased to locate the failure regions that are relatively close to the origin. Next, we limit the search area by defining cones that are formed by the failure samples. Cones are defined as structures whose apexes are considered to be the coordinate origin and whose bases are considered to be parts of the spherical surfaces that include the failure samples found in the previous step. Finally, the failure regions are refined by repeatedly bisecting the cone heights. The failure samples at the centers of the bases of the respective cones are used as the shift vectors for the mean-shift IS.

By determining the alternative distribution through the proposed method, we found that the subsequent yield estimation by IS became both stable and efficient. IS achieves multiple orders of reduction of the MC samples as compared with the conventional MC simulation. At a failure rate of 10^{-10}, which is equivalent to a 0.1% yield loss for a 10 Mb SRAM without redundancy, the number of MC samples required can be reduced by more than 10^6 times, which in turn will cause a speed-up of 10^6x.

The rest of this paper is organized as follows. In Sect. 2, we provide a background of this study, as well as present drawbacks of the existing methods. In Sect. 3, we propose a new method for determining the shift vector. In Sect. 4, we evaluate the effectiveness of the proposed method by estimating the failure rate of an SRAM cell. Finally, in Sect. 5, we present the conclusions of this study.

2. Monte Carlo Methods for Circuit Yield Analyses

In this section, the conventional MC method and the IS method are briefly reviewed.

2.1 Conventional Monte Carlo Method

The MC method is one of the most well-known statistical methods for estimating an expectation under a known distribution [4]. One advantage of this method is that it can be used to solve both nonlinear and linear problems, given the fact that it approximates the solution through a large number of simulations using randomly generated samples. Another advantage of this method is that it is flexible, and hence, can be applied to various problems for which no analytical solutions are available.

The principle of the MC method is explained below. Suppose we wish to calculate probability P that the value of a function f becomes less than a critical value f_0, i.e.,

$$f(\boldsymbol{x}) \le f_0. \tag{1}$$

Here, $f(\boldsymbol{x})$ is a function of an M-dimensional variable vector $\boldsymbol{x} = (x_1, ..., x_M)$, where $\boldsymbol{x}$ is a random variable that follows a probability distribution $p(\boldsymbol{x})$. The indicator function $I(\boldsymbol{x})$ is defined as

$$I(\boldsymbol{x}) = \begin{cases} 0, & \text{pass} \quad (f(\boldsymbol{x}) > f_0) \\ 1, & \text{failure} \quad (f(\boldsymbol{x}) \le f_0) \end{cases}. \tag{2}$$

We here assumed that the critical value gives an upper bound but more general bounds, such as to specify a range, can be considered in Eq. (2). Throughout the paper, sample $\boldsymbol{x}$ is referred to as the *pass sample* when $I(\boldsymbol{x}) = 0$. Otherwise, it is referred to as the *failure sample*, because in the context of failure rate estimation, the failure probability is evaluated. Further, the *failure region* and the *pass region* are defined as regions to which the failure samples and pass samples belong, respectively.

Probability P_{MC}, which is estimated using P by the conventional MC method, is calculated by

$$P_{\mathrm{MC}} = \frac{1}{N}\sum_{i=1}^{N} I(\boldsymbol{x}_i), \quad \boldsymbol{x}_i \sim p(\boldsymbol{x}). \tag{3}$$

Here, N is the number of MC trials. The MC sample $\boldsymbol{x_i}$ is generated to follow a probability density function $p(\boldsymbol{x})$. Convergence of the estimation of P_{MC} can be evaluated using its variance $\mathrm{Var}(P_{\mathrm{MC}})$ at the end of N-runs, which is given by the following equation,

$$\mathrm{Var}(P_{\mathrm{MC}}) = \frac{1}{N}\Big(\sum_{i=1}^{N} I(\boldsymbol{x}_i)^2 - {P_{\mathrm{MC}}}^2\Big). \tag{4}$$

For the estimation to be reliable, it is essential for variance $\mathrm{Var}(P_{\mathrm{MC}})$ to be small. The figure of merit for convergence, $\rho(P_{\mathrm{MC}})$, is defined as follows [9]:

$$\rho(P_{\mathrm{MC}}) = \frac{\sqrt{\mathrm{Var}(P_{\mathrm{MC}})}}{P_{\mathrm{MC}}}. \tag{5}$$

The figure of merit can be used to determine when the MC runs should be terminated. Assuming that estimation P_{MC} follows a Gaussian distribution, we find that the figure of merit at an accuracy of $100 \cdot (1-\epsilon)$ %, with a confidence of $100\cdot, (1-\delta)$ % is expressed as

$$\rho_0 = -\frac{\delta}{\Phi^{-1}(\epsilon/2)}. \tag{6}$$

Here, $\Phi^{-1}(p)$ is the inverse cumulative distribution function of the standard Gaussian distribution. Constants ϵ and δ are both close to zero. Equation (4) gives the required number of samples as

$$N_{\mathrm{MC}} = \frac{1 - P_{\mathrm{MC}}}{\rho^2 P_{\mathrm{MC}}} \simeq \frac{1}{\rho^2 P_{\mathrm{MC}}}. \tag{7}$$

We know that the required number of samples is inversely proportional to the failure probability. Hence, when the circuit yield is high, as in the case of SRAM cells, an intractably large number of samples and thus a long simulation time are required.

2.2 Importance Sampling

Importance sampling is a variance reduction technique used for MC simulations. The advantage of IS is that it reduces the number of MC trials [14]. In other words, it improves the reliability of estimation with a smaller number of MC samples.

P can be obtained by using an alternative probability distribution function $q(\boldsymbol{x})$ instead of the original distribution $p(\boldsymbol{x})$, but with some bias. Probability P_{IS}, which is estimated by the IS, is hence calculated by correcting the bias of using the alternative distribution as

$$P_{\mathrm{IS}} = \frac{1}{N} \sum_{i=1}^{N} I(\boldsymbol{x}_i) \cdot w(\boldsymbol{x}_i), \quad \boldsymbol{x}_i \sim q(\boldsymbol{x}). \tag{8}$$

Here, $w(\boldsymbol{x}) = p(\boldsymbol{x})/q(\boldsymbol{x})$ is the weight function that is used to adjust the bias. From Eqs. (4) and (8), the variance of P_{IS} is expressed as

$$\mathrm{Var}(P_{\mathrm{IS}}) = \frac{1}{N^2}\Big(\sum_{i=1}^{N} w(\boldsymbol{x}_i)^2 I(\boldsymbol{x}_i) - N {P_{\mathrm{IS}}}^2 \Big). \tag{9}$$

In this paper, $p(\boldsymbol{x})$ is assumed to be an M-dimensional joint Gaussian distribution with zero mean and standard deviations σ_j, such that

$$p(\boldsymbol{x}) = \prod_{j=1}^{M} \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left(-\frac{x_j^2}{2\sigma_j^2} \right). \tag{10}$$

We use the following Gaussian mixture distribution as the alternative probability distribution function for IS.

$$q(\boldsymbol{x}) = \sum_{i=1}^{N_C} m_i \cdot p(\boldsymbol{x} - \boldsymbol{r}_{\mathrm{IS}i}) \tag{11}$$

Here, m_i and $\boldsymbol{r}_{\mathrm{IS}i}$ are the mixing coefficients and the mean of the i-th Gaussian distribution, respectively. The mixing coefficients satisfy $\sum_{i=1}^{N_C} m_i = 1$. N_C is the number of Gaussian distributions that formulate $q(\boldsymbol{x})$.

2.3 Drawbacks of Mean-Shift IS

IS is an effective method that is used to accelerate rare event simulations. It is critical to determine a suitable alternative probability function $q(\boldsymbol{x})$ to make IS efficient, which is, in general, a very difficult task. It should be noted that when $q(\boldsymbol{x})$ is inappropriate, the efficiency and accuracy of IS will be deteriorated. In particular, mean-shift IS requires an appropriate choice of the mean-shift vector $\boldsymbol{s}$ [15]. As illustrated in the one-dimensional example shown in Fig. 1, it is desirable that $\boldsymbol{s}$ is located close to the pass/failure boundary that is nearest to the origin. This point is called as the *minimum norm point*. The samples around the minimum norm point have the most significant effect on the yield estimation, because these samples are more often observed among all other failure samples. It is found that with an increase in the sample dimension, the concentration of probability to the minimum norm point becomes more notable.

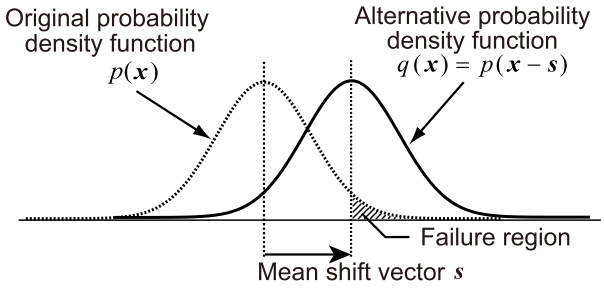

Fig. 1 Principle of importance sampling.

Methods to determine an appropriate mean-shift vector have been proposed in some literatures [8], [9]; however, the proposed methods have two major drawbacks. One drawback is that the minimum norm sample is determined by MC-runs using an M-dimensional uniform distribution. In this case, the search space becomes a hypercube, which is too large to determine an appropriate shift vector, particularly when the problem dimension is high. Hence, these methods tend to yield an ineffective shift vector [14]. The other drawback is that no approach has been provided to reduce the norm of the previously obtained minimum norm sample. This sample has the smallest norm within a set of generated failure samples. If we assume that the exploration space is large, the norm can be easily reduced by an additional effort. The coordinates around the small norm sample may also belong to the failure region and may have an even smaller norm. This hypothesis suggests that by carrying out additional searches around the samples that have small norm, it is possible to improve the quality of the candidate sample in terms of its norm. Hereafter, we refer to this process as refinement of minimum norm sample.

Without this refinement, it is uncertain whether or not an appropriate shift vector will be obtained, particularly when the problem dimension is high. Increasing the number of samples may help improve the quality of a suitable shift vector; however, this improvement comes at the cost of a longer runtime before the actual IS is begun. In light of these drawbacks, it is necessary to develop a structural method to improve the quality of the shift vector.

3. Hypersphere Sampling

This section describes the proposed hypersphere sampling that determines a suitable alternative probability function

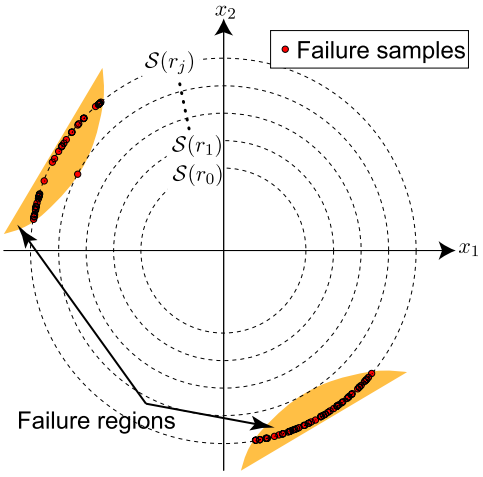

Fig. 2 Incremental hypersphere sampling.

$q(\boldsymbol{x})$. Hereafter, without loss of generality, we assume that variable $\boldsymbol{x}$ is normalized, i.e., $p(\boldsymbol{x})$ follows a standard joint Gaussian distribution with dimension M.

The proposed hypersphere sampling involves the following three steps:

1. Incremental hypersphere sampling
2. Failure sample clustering
3. Failure sample refinement.

Each step will be described in detail in the rest of this section.

3.1 Incremental Hypersphere Sampling

In order to locate the failure regions that are close to the coordinate origin, owing to which they significantly contribute to the failure probability, we first carry out incremental hypersphere sampling (IHS). The concept of IHS is illustrated in Fig. 2.

First, we randomly generate n_s samples on the surface of a hypersphere $\mathcal{S}(r_0)$ whose radius is r_0. Then, we carry out MC simulations using these n_s samples to check whether each sample falls in the failure region or not. If the number of failure samples found is less than a predetermined constant n_f, we increase the radius of the hypersphere to r_1, which is greater than r_0. For example, we uniformly increase the radius by one sigma, for each dimension, i.e., $r_1 = r_0 + 1$.

We generate n_s random samples on the surface of the new hypersphere $\mathcal{S}(r_1)$ and run simulations to obtain failure samples. The radius of the hypersphere, r_j, is increased until at least n_f failure samples are found on the surface of $\mathcal{S}(r_j)$.

The number of failure samples, n_f, with which we terminate the expansion of the hypersphere, are empirically determined. The average area-resolution to find a failure region is $\mathcal{S}(r_0)/n_s$. We should find all failure regions that affect the yield calculation. By setting $n_f = 10$, it is expected that a failure region that is larger than one tenth of the largest failure region shall be found.

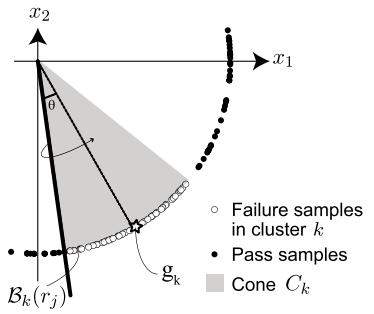

Fig. 3 Definition of cone C_k for failure sample cluster k.

3.2 Failure Sample Clustering

Let $\mathcal{S}(r_j)$ be the last hypersphere in the previous step, where a total of n_F ($\geq n_f$) failure samples are found. In this step, we divide the n_F-failure samples into clusters. This division is made to distinguish between failure regions so that the search range can be limited in the following refinement step. The furthest neighbor method [16], for example, can be used for this clustering where the cosine distance is the distance used for clustering. The cosine distance between two samples $\boldsymbol{s_1}$ and $\boldsymbol{s_2}$ is defined as follows:

$$\text{CosineDistance}(\boldsymbol{s_1}, \boldsymbol{s_2}) = 1 - \frac{\boldsymbol{s_1} \cdot \boldsymbol{s_2}}{|\boldsymbol{s_1}||\boldsymbol{s_2}|}. \tag{12}$$

A cluster is formed such that the largest distance between two samples in the cluster is less than 1.

We denote the center of gravity of a cluster k as $\boldsymbol{g}_k$. The pass sample closest to the center of gravity is denoted as $\boldsymbol{c}_k$. The angle formed by the two vectors $\boldsymbol{g}_k$ and $\boldsymbol{c}_k$, where both these vectors are considered to emanate from the origin, is given as

$$\theta_k = \arccos\left(\frac{\boldsymbol{g}_k \cdot \boldsymbol{c}_k}{|\boldsymbol{g}_k||\boldsymbol{c}_k|}\right). \tag{13}$$

A cone C_k for cluster k is obtained by rotating the half-line along $\boldsymbol{c}_k$ around vector $\boldsymbol{g}_k$, which is considered to be the axis of the cone. A two-dimensional example of the cone is illustrated in Fig. 3. The base of the cone is $\mathcal{B}_k(r)$, which is located on the surface of the sphere with radius r.

3.3 Failure Sample Refinement

We now intensively search the minimum norm sample of cluster k by bisection. The first bisection is illustrated in Fig. 4. Two bases of the cone are considered at radii $r_{\min}$ and $r_{\max}$. The third base $\mathcal{B}_k(r)$ is considered in the middle of the two bases, so that the samples are generated on it, and simulations are carried out to distinguish whether the samples are failure samples or not. Depending on whether at least one failure sample is found or not, the base of either $r_{\max}$ or $r_{\min}$ is replaced by that with the new radius $(r_{\max} + r_{\min})/2$ at the middle of the two bases. The following procedure will be used to efficiently search the minimum norm point for each

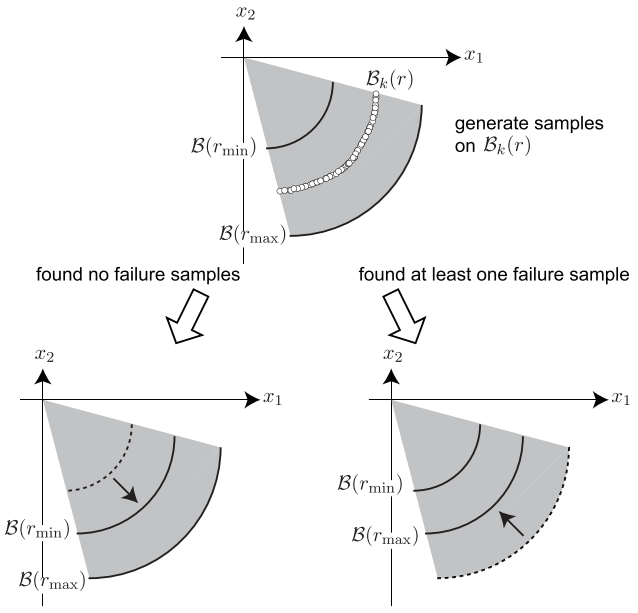

Fig. 4 Searching the minimum norm sample by bisection.

cluster.

1: $r_{\max} = r_1$
2: $r_{\min} = 0$
3: **repeat**
4: $r = (r_{\max} + r_{\min})/2$
5: S_{fail} := a set of failure samples found in $\mathcal{B}_k(r)$
6: **if** the number of $S_{\text{fail}} > 0$ **then**
7: $r_{\max} := r$
8: $F_k := S_{\text{fail}}$
9: **else**
10: $r_{\min} := r$
11: **end if**
12: **until** $r_{\max} - r_{\min} < r_{\text{th}}$
13: $\boldsymbol{g}'_k$:= the center of gravity of F_k
14: $\boldsymbol{r}_{\text{IS}k} := (r_{\max}/|\boldsymbol{g}'_k|) \cdot \boldsymbol{g}'_k$

Here, the number of samples in line 5 is $n_s \cdot (\theta_k/\pi)$. Bisection is terminated when the difference between $r_{\max}$ and $r_{\min}$ becomes smaller than a threshold r_{th} (line 12). We calculate the center of gravity on $\mathcal{B}_k(r_{\max})$ of the failure samples found in $\mathcal{B}_k(r_{\max})$ and define it as $\boldsymbol{r}_{\text{IS}k}$, which is the mean-shift vector of cluster k (lines 13, 14).

3.4 Failure Probability Estimation by Importance Sampling

Finally, we carry out IS using the Gaussian mixture distribution $q(\boldsymbol{x})$ defined by Eq. (11). The minimum norm sample $\boldsymbol{r}_{\text{IS}k}$ associated with cluster k is used as the mean-shift vector for the k-th Gaussian distribution in Eq. (11). The mixing coefficient m_k of the Gaussian distribution for cluster k is determined by the ratio of probabilities

$$m_k = \frac{p(\boldsymbol{r}_{\text{IS}k})}{\sum_{k=1}^{N_c} p(\boldsymbol{r}_{\text{IS}k})}. \tag{14}$$

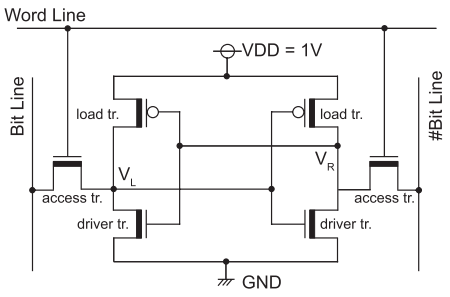

Fig. 5 Schematic of a 6-transistor SRAM cell.

IS terminates when the figure of merit of convergence, $\rho(P_i)$, for the estimation becomes less than ρ_0.

4. Numerical Experiments

In this section, we evaluate the accuracy and runtime of IS using the proposed method.

The circuit used to estimate the failure probability is a six-transistor SRAM cell shown in Fig. 5. The threshold voltage (Vth), gate length (L_{g}), carrier mobility (μ), and gate oxide thickness (T_{ox}) of all the transistors are considered as variables. The 65-nm predictive technology model (PTM) [17] is used as a transistor model. The parameters of each transistor are assumed to follow joint Gaussian distributions whose means and standard deviations are summarized in Table 1.

Pass or failure of an SRAM cell is determined using a signal noise margin (SNM). A sample is considered to be a failure sample when the SNM of the SRAM cell is 0 or less. The failure probabilities of

- a 6-dimensional (6-D) problem that considers Vth variations of all transistors;
- a 12-dimensional (12-D) problem that considers Vth and L_{g} variations of all transistors;
- an 18-dimensional (18-D) problem that considers Vth, L_{g}, and μ variations of all transistors; and
- a 24-dimensional (24-D) problem that considers Vth, L_{g}, μ, and T_{ox} variations of all transistors

are estimated using the proposed technique and the conventional MC method.

Estimations obtained with the proposed method, P_{IS}, should match with those obtained with the conventional MC method, P_{MC}, although P_{MC} may include some error. The termination criterion ρ_0 is set to 0.1, for both the proposed IS and the conventional MC method. This implies that the both estimations have errors less than ±20%, with 95% confidence. A comparison between the runtimes of the proposed IS and the conventional MC method is made on the basis of the number of circuit simulations required to achieve a termination criterion, because the runtime is dominated by a non-linear DC analysis [18] that is used to calculate the SNM. In the case of the proposed method, IHS starts with an initial radius of 3-sigma ($r_0 = 3$) and the number of samples, n_s, for the 6-D, 12-D, 18-D, and 24-D problems are 5×10^3,

Table 1 Variability parameters of transistors.

Transistors	Access		Driver		Load	
	mean	SD	mean		mean	SD
Threshold voltage (mV)	-	18.0	-	22.0	-	30.0
Gate length (nm)	65.	1.	65.	1.	65.	1.
Mobility (mm^2/Vs)	4.91	0.50	4.91	0.50	0.574	0.054
Gate oxide thickness (nm)	1.85	0.04	1.85	0.04	1.950	0.042
Gate width (nm)	110	-	120	-	80	-

Table 2 Comparison of estimated failure rate and the number of required samples between the conventional MC method and the proposed IS. The proposed method shows the ranges for 20 runs with different random seeds.

	MC method		Proposed method						
	P_{MC}	# of samples	P_{IS}			# of samples ($\times 10^3$)			
dim.		($\times 10^3$)	Min	Max	Median	IHS	Bisection	IS	Total
6	-	(1.8e+07)	4.3e-09	6.5e-09	5.6e-09	25.0	11.8–13.1	0.8–2.0	38.0–39.9
12	–	(3.1e+05)	2.8e-07	3.9e-07	3.3e-07	50.0	29.0–33.2	0.6–2.9	80.1–84.2
18	1.2e-06	8.7e+04	7.9e-07	1.5e-06	1.2e-06	75.0	31.2–52.2	1.4–20.6	122.1–140.5
24	1.5e-06	6.6e+04	1.2e-06	1.7e-06	1.5e-06	120.0	75.1–114.2	1.3–56.7	197.9–251.8

1×10^4, 1.5×10^4, and 2×10^4, respectively. n_s is determined such that it is directly proportional to the number of dimensions, M. Ideally, n_s should be increased in proportion to 2^M, but it is unfeasible when M becomes large. Hence n_s is determined empirically. In our experiments, failure region becomes large as we enlarge the radius of the hypersphere, and thus linear increase of n_s would find failure regions successfully. Determination of appropriate number of samples for more general examples is one of our future work.

4.1 Accuracy Evaluation

Table 2 summarizes the results of the comparison of the failure-rate estimations and the numbers of samples used for the estimations between the conventional MC method and the proposed IS. The proposed IS is repeated 20 times to confirm the stability of the estimation. The maximum and the minimum number of samples required in each run are listed in this table. In low-failure probability problems, the MC method requires a very long runtime. The numbers of samples required for the 6-D and 12-D problems using the conventional MC method are the estimates obtained using Eq. (7); thus, they are listed in parenthesis. It takes 9 days to obtain 3.1×10^8 samples, under the assumption that the simulation takes 2.5 ms for each sample.

Considering the estimations obtained by the conventional MC method as reference estimations, we find that the estimation errors of the proposed method are within 20% for more than 95% trials. For the 6-D and 12-D problems, the respective averages obtained using the proposed method are used as the accuracy references. We compared the proposed method and the conventional MC method by their transistor parameters in the failure regions for the 24-D problem. Only a few failure samples could be obtained by the conventional MC method, but obtained transistor parameters are nearly identical in both methods.

The number of Gaussian distributions, N_c in Eq. (11), is determined by the result of clustering. The number of clusters should be closely related to the number of failure regions. In most problems, two clusters were formed, which we think is reasonable considering that the 6-transistor SRAM cell is symmetric. Only in two trials in the 24-D problem, three clusters were formed. More than one cluster may be required to represent a failure region, particularly when the problem dimension is high. This is the reason why three clusters were formed only in the 24-D problems. Unlike the case in which one or more failure regions are not covered with any cluster, in the case in which a failure region is covered by two or more clusters, the estimation accuracy is unaffected. As long as clusters are formed in all the failure regions, samples will be generated in all failure regions, later in the IS step.

On the other hand, if a small number of clusters is found in the repetitive runs of the proposed method on the same problem, it may indicate that a few failure regions are missing, which will lead to underestimation of the failure rate. For example, the estimation becomes approximately half if one out of the two failure regions is not considered. Therefore, to avoid this issue, it is essential to not set the value of n_f too small. Throughout the experiments, $n_f = 10$ is used.

The furthest neighbor method [16] has been chosen as the clustering algorithm in our implementation, because it is more suitable than the nearest neighbor method that is more widely used. The nearest neighbor method tends to choose the newly constructed cluster as the candidate cluster for merging [16], leading to the formation of large clusters between the two failure regions. It would be necessary to optimize the standard deviation of the alternative distribution if we adopt the nearest neighbor method; otherwise, IS would become inefficient.

4.2 Effect of Standard Deviations

The mean-shift method was used in the previous evaluations. During the evaluations, standard deviations of the

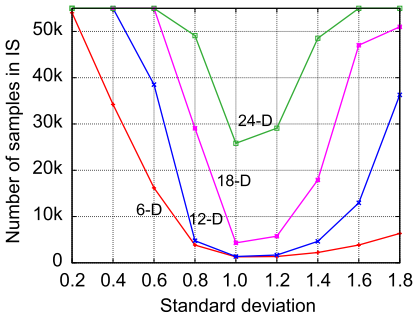

Fig. 6 Number of samples required in importance sampling as functions of the standard deviations of Gaussian distributions in the alternative distribution.

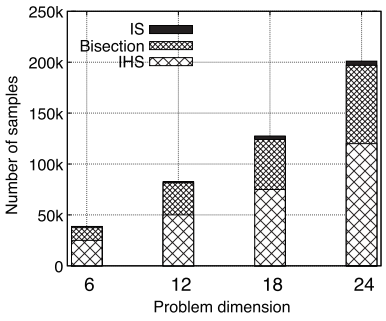

Fig. 7 Breakdown of the number of samples.

joint Gaussian distributions are set equal to the original standard deviations. In this subsection, the standard deviation is also changed.

Figure 6 shows the number of samples required in IS as functions of the standard deviation of the alternative distribution. The points shown in this figure indicate the averages of the 20 trials. For all dimensions, the number of samples becomes the smallest when the standard deviation is around 1.0–1.2. In the 6-D problem, the number of samples is relatively insensitive to the standard deviation. With an increase in the problem dimension, the range over which a suitable standard deviation can be chosen becomes narrower. Although the original standard deviation of 1.0 seems appropriate for all dimensions between 6- and 24-D problems in this example, it may be problem- as well as cluster-size-dependent. In future, we intend to automate the determination of the appropriate standard deviation for a high-dimensional problem.

4.3 Effect of Problem Dimensions

According to Table 2, the proposed method reduces the required number of samples by 2–5 orders of magnitude as compared with the conventional MC method. Because the runtime is directly proportional to the number of samples, the proposed method is significantly faster than the conventional MC method. Further, the proposed method becomes even faster when the failure probabilities become low.

Figure 7 shows the breakdown of the samples in each step. The numbers shown in this figure denote the medians of the 20 trials. The required sample numbers increase with an increase in the dimension of the problem. Given the fact that the search space increases with the problem dimension, it becomes increasingly difficult to determine the failure samples. The number of samples should be increased exponentially such as in the order of 2^M to search for the minimum norm samples with equal density; however, it is not feasible to do so when M becomes large. Thus, in this evaluation, we increased the number of samples linearly with the problem dimension.

It should be noted that only a few samples are required for IS as compared with those required for the preparation steps to obtain the shift vectors. IS converges very quickly as long as a suitable alternative distribution is determined. This result validates the importance of the choice of shift vectors. The proposed method found appropriate shift vectors for each problem dimension.

The number of samples required by the proposed IS increases with an increase in the problem dimension. The speed-up ratio of the proposed method over the conventional MC decreases accordingly. This result does not necessarily indicate that IS is ineffective for high-dimensional problems; rather, the effectiveness of IS becomes less prominent as compared with the conventional MC method as the failure probability becomes high. In these experiments, the failure probability increases with an increase in the number of variation sources. This is completely understandable because variation sources are incrementally added with an increase in the problem dimension. In this case, according to Eq. (7), the number of samples required by the conventional MC method decreases with an increase in the problem dimension, because the required sample is inversely proportional to the failure probability. On the other hand, thorough search for the minimum norm sample requires exponentially longer time as the problem dimension increases. These results enabled the runtimes of the conventional MC and the proposed method to be approximately equal. However, according to Table 2, IS is still more than 10x more efficient than the conventional MC method. The problems of even higher failure probability than the ones listed in Table 2 can be and should be analyzed with the conventional MC method. If the failure probability is maintained to be sufficiently low in high-dimensional problems, such as those in the order of 10^{-9}, the proposed method is more efficient than the conventional MC method.

The difficulty of analyzing high-dimensional problems should be carefully studied to maximize efficiency. Ideally, according to Eq. (8), weight $w(\boldsymbol{x})$ should be constant for any $\boldsymbol{x}$. However, this constant weight can only be achieved when ratio $p(\boldsymbol{x})/q(\boldsymbol{x})$ is constant. For this ratio to be constant, it is essential for the shape of the alternative distribution to be in proportion to the failure distribution; however, it is difficult to meet this requirement with an increase in the problem dimension. In order to confirm this, we calculated the standard deviation of ratio $p(\boldsymbol{x}_f)/q(\boldsymbol{x}_f)$ and normalized it by its average. The results of this calculation are 2.7–3.2, 2.8–4.6,

Table 3 Runtime comparison with other importance sampling literatures.

		# of samples		
	Estimation p	Preprocess	IS	Total
Proposed	5.6e-9	37.5k	1150	38.6k
[19]	5.6e-9	75.3k	780	76.1k
Sequential-IS [11]	3.2e-9	0	11k.0	11.0k
Variance expansion [20]	5.7e-9	0	296k	296k
Norm minimization [9]	5.0e-9	10.0k	220k	230k
Gibbs sampling [12]	3.3e-9	21.8k	930	22.7k
Gaussian approximation	4.1e-9	4k	-	4k

2.9–11.3, and 4.1–34.5 for 6-D, 12-D, 18-D, and 24-D problems, respectively. It is found that the variation in the ratio increases with the problem dimension. Accordingly, an increase in the calculation time with the problem dimension is unavoidable. When the calculation time becomes too large, it is preferred to reduce the problem dimension by eliminating low-sensitivity variables.

From Fig. 7, it is observed that the number of samples required for the M-dimensional analysis is approximated as $(3.2 \times 10^4) \cdot 1.08^M$ in this example. This implies that 90-dimensional problems with a similar probability problem can be estimated within a day.

4.4 Comparison with Other Work

The proposed method is compared with other techniques that also use mean-shift IS. The failure probability of an SRAM cell is estimated using the techniques described in [9], [19], [20]. In this evaluation, the 6-D problem is considered. In Table 3, the estimation results and the required number of samples are compared. The medians of 20 trials are listed for the proposed method, [19] and [11]. For the other techniques, the medians of 3 trials are listed. The reason why we compare medians rather than means is that means are easily influenced by outliers.

In Table 3, the column "preprocess" lists the number of samples used for shift-vector determination process(es) before starting IS. Further, the column "IS" lists the required number of samples in IS until the equal convergence criterion is achieved. The four methods, i.e., the proposed method, the method in [19], sequential-IS [11], the variance expansion method [20], the norm minimization method [9], and Gibbs sampling method [12], estimated similar failure probabilities; however, the total number of samples for these methods differed by an order of magnitude except for the sequential-IS and Gibbs sampling methods. However, in our experiment, these methods underestimated the failure probability. Also as a comparison, we estimate a failure probability assuming that SNM follows a Gaussian distribution. To determine the mean and variance of the Gaussian distribution, 4000 random samples are used. Although the median of the estimates looks relatively good, the estimates vary widely from 2.3e-9 to 8.1e-9 in 20 trials. Furthermore, we can not control this variation using ρ_0 unlike IS.

A method in [19] is a previous version of the proposed method. The differences between the proposed method and the one in [19] are 1) [19] uses the region between concentric hyperspheres in IHS and 2) [19] uses decremental sampling instead of the bisection method. The proposed method showed an improvement in these two points, achieving about twice the efficiency as that of [19] in the step of shift-vector determination.

Sequential-IS [11] conducts IS while determining failure distribution using particle filter. This method does not require preprocess, and hence its calculation time is the shortest among the six methods. However, the particle filter misses to cover one of the existing failure regions. This is why the particle filter estimates the failure probability about half of those obtained by other methods because there are two failure regions of equal size in this example.

The variance expansion method [20] uses alternative distribution with three times larger standard deviation than the original standard deviation. Further, this method does not require to search the shift vector; however, the convergence of IS is slower than that of the proposed method.

Norm minimization [9] searches the nearest failure sample to the coordinate origin using uniform distribution. The range of $[-10\sigma, 10\sigma]$ is used for the range of 6-D joint uniform distribution. For the sake of fair comparison, two Gaussian distributions whose mean-shift vectors are the two nearest failure samples to the coordinate origin are used for IS.

Gibbs sampling method [12] firstly generates Gibbs samples whose distribution is equal to $p(\boldsymbol{x})$. And then, it calculates mean and covariance of these samples to define a multivariate normal distribution as an alternative probability distribution function for IS. This method uses only one normal distribution, which caused the underestimation of the failure probability, by the same reason as the case of the particle filter.

5. Conclusion

This paper proposes a method termed hypersphere sampling. The proposed method efficiently determines appropriate shift vectors that are critically important in mean-shift importance sampling. With the proposed method, the yield estimations of high-dimensional and low-failure probability circuits are significantly accelerated. The results of the experiment performed on an SRAM circuit verified the effectiveness of the proposed method, reducing the number of required Monte Carlo simulation runs by 2–5 orders of magnitude as compared to a conventional MC method.

Acknowledgment

This work was partially supported by JEDAT, Special Coordination Funds for Promoting Science and Technology and NEDO, and VLSI Design and Education Center (VDEC) in collaboration with Synopsys, Inc.

References

[1] S. Nassif, "Design for variability in DSM technologies," Proc. IEEE

International Symposium on Quality Electronic Design, pp.451–454, 2000.
[2] B. Cheng, S. Roy, and A. Asenov, "The impact of random doping effects on CMOS SRAM cell," Proc. European Solid-State Circuits Conference, pp.219–222, 2004.
[3] K. Agarwal and S. Nassif, "The impact of random device variation on SRAM cell stability in sub-90-nm CMOS technologies," IEEE Trans. Very Large Scale Integr. (VLSI) Syst., vol.16, no.1, pp.86–97, 2008.
[4] W.G. Cochran, Sampling techniques, Wiley, New York, 1977.
[5] A. Singhee, J. Wang, B. Calhoun, and R. Rutenbar, "Recursive statistical blockade: An enhanced technique for rare event simulation with application to SRAM circuit design," Symposium on VLSI Technology, Digest of Technical Papers, pp.131–136, 2008.
[6] T. Doorn, E. ter Maten, J. Croon, A. Di Bucchianico, and O. Wittich, "Importance sampling monte carlo simulations for accurate estimation of SRAM yield," Proc. European Solid-State Circuits Conference, pp.230–233, Sept. 2008.
[7] G. Chen, D. Blaauw, T. Mudge, D. Sylvester, and N.S. Kim, "Yield-driven near-threshold SRAM design," Proc. IEEE/ACM International Conference on Computer-Aided Design, pp.660–666, 2007.
[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," Proc. IEEE/ACM Design Automation Conference, pp.69–72, 2006.
[9] L. Dolecek, M. Qazi, D. Shah, and A. Chandrakasan, "Breaking the simulation barrier: SRAM evaluation through norm minimization," Proc. IEEE/ACM International Conference on Computer-Aided Design, pp.322–329, 2008.
[10] A.J. Mcneil, D. Mathematlk, and E. Zentrum, "Estimating the tails of loss severity distributions using extreme value theory," ASTIN Bulletin, vol.27, no.1, pp.117–137, 1997.
[11] K. Katayama, S. Hagiwara, H. Tsutsui, H. Ochi, and T. Sato, "Sequential importance sampling for low-probability and high-dimensional SRAM yield analysis," Proc. IEEE/ACM International Conference on Computer-Aided Design, pp.703–708, IEEE, 2010.
[12] C. Dong and X. Li, "Efficient SRAM failure rate prediction via Gibbs sampling," Proc. IEEE/ACM Design Automation Conference, pp.200–205, 2011.
[13] A. Dembo and O. Zeitouni, Large Deviations Techniques and Applications, Second ed., Springer, 1998.
[14] T.C. Hesterberg, Advances in Importance Sampling, Ph.D. Thesis, Statistics Department, Stanford Univ., 1988.
[15] G. Schueller, H. Pradlewarter, and P.S. Koutsourelakis, "A comparative study of reliability estimation procedures for high dimensions," ASCE Engineering Mechanics Conference, 2003.
[16] T. Kamishima, "A survey of recent clustering methods for data mining (part 2)," J. Japanese Society for Artificial Intelligence, vol.18, no.1, 2003.
[17] W. Zhao and Y. Cao, "New generation of predictive technology model for sub-45 nm early design exploration," IEEE Trans. Electron Devices, vol.53, pp.2816–2823, Nov. 2006.
[18] L.W. Nagel, SPICE2: A Computer Program to Simulate Semiconductor Circuits, Ph.D. Thesis, EECS Department, University of California, Berkeley, 1975.
[19] T. Date, S. Hagiwara, K. Masu, and T. Sato, "An efficient technique to search failure-areas for yield estimation via partial hypersphere," IEICE Technical Report, VLD2009-105, March 2010.
[20] T. Doorn, E. ter Maten, J. Croon, A. Di Bucchianico, and O. Wittich, "Importance sampling monte carlo simulations for accurate estimation of SRAM yield," Proc. European Solid-State Circuits Conference, pp.230–233, 2008.

Shiho Hagiwara received B.E., M.E., and Ph.D. degrees from the Tokyo Institute of Technology, Japan, in 2006, 2008, and 2011, respectively. In 2011, she joined Fujitsu Laboratories Ltd., Kawasaki, Japan. Her research interests include CAD algorithms for VLSI design, design for manufacturing and power integrity.

Takanori Date received a B.E. degree from the Shibaura Institute of Technology, Tokyo, Japan, in 2008, and a M.E. degree from the Tokyo Institute of Technology, Tokyo, Japan, 2010. In 2010, he joined Oki Electric Industry Co., Ltd., Tokyo, Japan.

Kazuya Masu received B.E., M.E., and Ph.D. degrees in Electronics Engineering from the Tokyo Institute of Technology, Tokyo, Japan, in 1977, 1979, and 1982, respectively. He was with the Research Institute of Electrical Communication, Tohoku University, Sendai, Japan, since 1982. Since 2000, he has been with the Precision and Intelligence Laboratory, Tokyo Institute of Technology, Yokohama, Japan, and is currently a professor at Solutions Research Laboratory, Tokyo Institute of Technology, Yokohama, Japan. He also serves Director, ICE Cube Center, Tokyo Institute of Technology. He was a visiting professor at the Georgia Institute of Technology in 2002 and 2005. His current interests include signal integrity and gigahertz signal propagation in a multilevel interconnection of Si ULSI, scalable and reconfigurable RF CMOS circuit technology, design and implementation for integration of diverse functionalities on CMOS. He is a member of the IEEE, the Japan Society of Applied Physics (JSAP), the Institute of Electrical Engineers of Japan (IEEJ), and the institute of Electronics, Information and Communication Engineers (IEICE). He was awarded as Fellow of JSAP and IEEJ.

Takashi Sato (M'98) received B.E. and M.E. degrees from Waseda University, Tokyo, Japan, and a Ph.D. degree from Kyoto University, Kyoto, Japan. He was with Hitachi, Ltd., Tokyo, Japan, from 1991 to 2003, with Renesas Technology Corp., Tokyo, Japan, from 2003 to 2006, and with the Tokyo Institute of Technology, Yokohama, Japan. In 2009, he joined the Graduate School of Informatics, Kyoto University, Kyoto, Japan, where he is currently a professor. He was a visiting industrial fellow at the University of California, Berkeley, from 1998 to 1999. His research interests include CAD for nanometer-scale LSI design, fabrication-aware design methodology, and performance optimization for variation tolerance. Dr. Sato is a member of the IEEE and the Institute of Electronics, Information and Communication Engineers (IEICE). He received the Beatrice Winner Award at ISSCC 2000 and the Best Paper Award at ISQED 2003.