Statistical Leakage Analysis Using Gaussian Mixture Model

Hyunjeong Kwon¹, Student Member, IEEE, Mingyu Woo², Student Member, IEEE, Young Hwan Kim¹, Senior Member, IEEE, and Seokhyeong Kang¹, Member, IEEE.
¹Electrical Engineering, Pohang University of Science and Technology, 77 Cheongam-Ro. Nam-Gu, Pohang, Gyeongbuk, South Korea, 37673
²Computer Engineering, University of California San Diego, La Jolla, CA 92093, USA

Corresponding author: Seokhyeong Kang (e-mail: shkang@postech.ac.kr).

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the “ICT Consilience Creative program” (IITP-2018-2011-1-00783) supervised by the IITP(Institute for Information & communications Technology Promotion).

ABSTRACT In the design process of advanced semiconductor devices, statistical leakage analysis has emerged as a major step due to uncertainties in the leakage current caused by the process variations. In this paper, a novel statistical leakage analysis which uses Gaussian mixture model (GMM) as the density function of leakage current is proposed. To estimate the probability density function, our proposed method clusters the rapidly converged leakage data using the GMM. The GMM can represent any distributions, so it is suitable to estimate the leakage distribution, which varies as the technology node or operating condition changes. In addition, our proposed method (SLA-GMM) defines a terminating condition that guarantees the convergence of the leakage data and prevents the under fitting or over fitting in the GMM modeling process. With sequential addition, SLA-GMM significantly reduced the error that can occur during the addition process. In studies with a goodness-of-fit test, SLA-GMM achieved up to 98% and 94% improvements in the Chi-square static and the K-S static compared to the previous method based on an analytic model.

INDEX TERMS Expectation-Maximization Algorithm, Gaussian Mixture Model, Machine Learning, Statistical Leakage Analysis.

I. INTRODUCTION

Low power usage is mobile computing devices. To realize various strategies to reduce power use, the power must be accurately estimated during the design stage. Especially, leakage power is increasingly dominant in many applications [1-4]. Meanwhile, the process variations increase as leakage size is scaled down [5-7]. Process variations lead to variations in leakage current $I_L$ because it is directly dependent on process parameters. For this reason, variation-aware $I_L$ analysis, called statistical leakage analysis (SLA), has become an important step in the process of designing low-power devices. The previous SLA researches can be classified into two main approaches; sampling-based methods and analytic model-based methods.

Sampling-based methods use extracted parameter samples to simulate $I_L$. To reduce the complexity of traditional Monte-Carlo (MC) simulation, efficient leakage models have been developed [8, 9] and the convergence time of sampling has been reduced [10-12]. Sampling-based methods store all leakage data or histograms to estimate the leakage distribution.

Analytic model-based methods use an analytic model to represent $I_L$. This method does not require numerous simulations, so it can generate feasible models quickly. The first-order model [17] is the most representative of these methods. The first-order model is based on the assumptions that the log($I_L$) has linear relationships with process parameters, and
that the leakage follows a lognormal distribution. In advanced technology nodes, the generalized extreme value (GEV) distribution is used to improve the accuracy [18]. A proposed fourth-order model [19] is based on the assumption that the leakage distribution follows the GEV distribution.

The shape of leakage distribution varies as the technology node advances, and as operating conditions such as supply voltage change. Therefore, whenever the technology or operating condition changes, a new model must be developed to fit the leakage distribution to a well-known continuous function. For example, the accuracy of the first-order model [17] is significantly reduced in technologies that are more recent than BSIM4 models [20] due to their non-lognormality characteristics [21]. Furthermore, the GEV also cannot adapt to changes in the shape of the leakage distribution.

SLA cannot readily sum the leakage distributions of small circuits. The summation allows the leakage analysis at a smaller level than the full-chip level. Basically, the distribution of the sum of two arbitrary distributions can be obtained using the convolution of the two distributions. However, it is inefficient to perform a convolution every time the summation is performed and the convolution data points are changed. To avoid this, if the approximation to represent that the summation result using the same density function [18, 22, 23] is applied, the errors occurred in the summation process seriously degrade the accuracy.

Here, we propose a novel SLA method that accurately estimates the chip leakage distribution by using a Gaussian Mixture Model (GMM) (Fig. 1). To estimate the leakage distribution of a circuit, our proposed method (SLA-GMM) can represent any arbitrary function without fixing the shape of the leakage distribution to a particular continuous function. The terminating condition of SLA-GMM is defined using the error bound of Gaussian function. Finally, the summation in SLA-GMM does not require a convolution process, because the summation result can be represented as another GMM by using only the parameters of GMMs without any approximation.

The contributions of SLA-GMM are:

- It can represent any leakage distribution to maintain the accuracy even when the technology node or the operating point changes.

\[ p(x) = \sum_{i} \omega_i \cdot N(x | \mu_i, \sigma_i) = \sum_{i} \frac{1}{\sigma_i \sqrt{2\pi}} e^{-\frac{(x-\mu_i)^2}{2\sigma_i^2}} \]

where \( \mu_i \) is the mean, \( \sigma_j \) is the standard deviation, and \( \omega_i \) is the weight of the \( i^{th} \) Gaussian component. Because the GMM consists of several Gaussian functions, it can represent various classes of continuous functions with reasonable accuracy. Hence the GMM can be used to estimate the probability density function (PDF) of the observed data. To utilize the GMM to estimate the distribution, the data must be clustered with several Gaussian components. The parameters of GMM such as \( N_{GMM}, \mu_i, \sigma_j, \) and \( \omega_i \) are determined during the clustering process. The most frequently-used method for GMM clustering is the expectation-maximization (EM) algorithm.
B. Expectation-Maximization Algorithm

EM algorithm iteratively finds the maximum likelihood estimates of the parameters of statistical models. It performs an E-step and an M-step iteratively. In the E-step, a function for likelihood is created using current parameter estimates. In the M-step, the parameters are updated by maximizing the likelihood function generated in the E-step. If the EM algorithm is used for clustering in GMM modeling, the following E- and M-steps are performed iteratively.

1) **E-step:** For each data point and Gaussian component, a membership weight \( \omega_{ik} \) is calculated. \( \omega_{ik} \) is the probability that the \( i \)th data point belongs to the \( k \)th cluster \( C_k \). When \( N \) is the number of data and \( K \) is the number of clusters, the result of this step is an \( N \times K \) matrix of membership weights where the elements sum to one in each row.

2) **M-step:** Using the data and the matrix of membership weights, a new parameter set is determined as (2).

\[
\begin{align*}
\omega_{ik}^{\text{new}} &= \frac{N_k}{N} \\
\mu_k^{\text{new}} &= \frac{1}{N_k} \sum_{i=1}^{N} \omega_{ik} x_i \\
\sigma_k^{2,\text{new}} &= \frac{1}{N_k} \sum_{i=1}^{N} \omega_{ik} (x_i - \mu_k^{\text{new}})^2
\end{align*}
\]

where \( N_k = \sum_{i=1}^{N} \omega_{ik} \) is the effective number of data that are included in the \( k \)th Gaussian component.

III. STATISTICAL LEAKAGE ESTIMATION USING GAUSSIAN MIXTURE MODEL

SLA-GMM estimates the leakage distribution of sub-block circuits and adds more than two leakage distributions sequentially. SLA-GMM consists of two main parts (Fig. 3). The first part is GMM modeling to estimate the leakage distribution. This step generates the leakage data for the GMM modeling, then after clustering, calculates the required number of leakage data. If the current number of data is less than the required number to ensure convergence, then additional leakage data are generated. The second part is sequential addition, which is essential in the SLA and helps to improve the scalability of SLA-GMM. SLA-GMM adds multiple leakage distributions sequentially, using the GMM parameters. This step can be useful to obtain the leakage distribution of a huge system that consists of many sub-blocks.

A. ESTIMATION OF THE LEAKAGE DISTRIBUTION USING GMM

In the SLA-GMM, we use three steps to estimate the leakage distribution. (1) We utilize the rapidly-converged samples to reduce the required number of simulation data. To achieve this goal, we use a Sobol sequence [24] for parameter sampling instead of pseudorandom numbers. (2) We use the GMM as the leakage distribution model to compensate for the loss of accuracy caused by use of the reduced number of simulation data. (3) We use a new method that calculates the number of simulation data required for convergence, and dynamically supplement the additional data if necessary. The
can be determined is the allowed percentage error, and, the mean-
and standard deviation is the confi-
as a rule of thumb [27] 

2) CLUSTER VALIDATION

The clustering result is first validated using cluster cohesion, which evaluates the tightness of the cluster of leakage data; the measure is the sum of squares within cluster (SSW)

$$SSW = \sum_i \sum_{x \in C_i} (x - m_i)^2$$

where \( x \) is a datum in the \( i \)th cluster \( C_i \) and \( m_i \) is the mean of the data in \( C_i \).

Then the separation of clusters is calculated as the sum of squares between clusters (SSB)

$$SSB = \sum_i |C_i| (m - m_i)^2$$

where \( |C_i| \) is the number of data in \( C_i \) and \( m \) is the mean of all data.

To optimize the number \( \kappa \) of clusters, we used the WB index [25].

$$WB\text{-index} = \kappa \cdot \frac{SSW}{SSB}$$

The index is calculated sequentially. Usually, \( \kappa \) is initialized to two [26], then increased and WB-index is calculated each time. After the maximum \( \kappa \left( \sqrt{L/2} \right. \) as a rule of thumb [27]) is reached, \( \kappa \) that fields the smallest \( WB\text{-index} \) is selected as the optimal \( \kappa \).

3) CALCULATION OF THE REQUIRED NUMBER OF DATA

To define the terminating condition and to prevent underfitting or overfitting, the number of required data should be calculated. The calculation for the convergence is simplified when the data are extracted from Gaussian distributions. SLA-GMM defines the required number of leakage data by applying the error for data obtained from Gaussian distribution as follows.

The required number of data from a Gaussian distribution with mean \( \bar{x} \) and standard deviation \( S_x \) can be determined as [28]

$$N = \left[ \frac{100z_cS_x}{Ex} \right]^2$$

where \( E \) is the allowed percentage error, \( z_c \) is the confidence level coefficient. If \( z_c = 1.96 \) and \( E = 1 \), the meaning of (3) is that using \( N \) data, we are 95% confident that the Gaussian distribution fitted from the data does not differ by more than 1% from the real Gaussian distribution.

FIGURE 5. 256 samples transformed from (a) the pseudo random numbers and (b) the Sobol sequence.

calculated number of simulations can be used as a terminating condition and is useful to prevent overfitting or underfitting when estimating the parameters of the GMM.

The estimation steps of the leakage distribution consist of three processes (Fig. 4). (1) Leakage data are generated for the GMM modeling, and clustering is performed. (2) The clustering is validated using a certain index. (3) After optimal clustering is performed, the required number of data is calculated and compared with the current number of data. If the current number of data is insufficient, leakage data for modeling are added. Details follow.

1) GENERATION OF LEAKAGE DATA FOR GMM MODELING

To generate leakage data to be used for the GMM modeling, parameter samples for the leakage simulation are extracted from the parameter spaces. The process parameters are assumed to follow Gaussian distributions. Therefore, after uniformly-distributed samples are extracted, we use the Box-Muller method to transform them to normally-distributed samples.

The purpose of SLA-GMM is to reduce the number of MC runs for efficiency. Therefore, the parameter samples should be extracted with a high convergence rate. SLA-GMM uses the Sobol quasi random sequence [24] rather than pseudorandom numbers as the uniform samples. The Sobol sequence converges faster than the pseudo random numbers because it is determined by considering the discrepancy, which is a mathematical quantity that represents the non-uniformity of the points. Transformation using the Sobol sequence is closer to Gaussian than is transformation using pseudorandom numbers (Fig. 5).

Clustering is performed on leakage data obtained using any leakage model and the parameter samples. Then the quality of clustering is evaluated in two ways. (1) Cluster validation is performed to evaluate the degree of density and separation of data. (2) The required number of data is calculated. If the required number of data is larger than the current number of data, additional data are collected. Details follow.
Each component of the GMM is a Gaussian distribution and its clustered data, so each cluster must have more data than \(N\) in (3). Therefore, after leakage data are obtained and GMM modeling is completed with optimal \(\kappa\); we test whether the numbers of data in all clusters satisfy (3). If this condition is satisfied, the data of each cluster can be regarded with confidence level to come from a Gaussian distribution. However, if one of the clusters does not satisfy the condition, additional data for GMM modeling are added. This step exploits the increase in accuracy of estimating the leakage distribution. Generally, the modeling errors before the summation operation are accumulated as summation errors. This error is exacerbated when approximations must be used to allow sequential addition. The sum of two distributions represented in the GMMs can be also expressed in the GMM without any approximation, so SLA-GMM extremely reduces the rate of error accumulation during the summation step.

For a Gaussian distribution, \(f_z(z)\) can be simply summarized after (7) is solved. The complexity of this calculation is greatly reduced using (8).

\[
X : N \sim (\mu_1, \sigma_1^2), Y : N \sim (\mu_2, \sigma_2^2) \\
\rightarrow Z : N \sim (\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2) \tag{8}
\]

SLA-GMM uses GMM as the leakage distribution, so

\[
X : p_1 = \sum_{j} \omega_j \frac{1}{\sigma_j \sqrt{2\pi}} e^{-\frac{(z-\mu_j)^2}{2\sigma_j^2}} \\
Y : p_2 = \sum_{k} \omega_k \frac{1}{\sigma_k \sqrt{2\pi}} e^{-\frac{(z-\mu_k)^2}{2\sigma_k^2}} \tag{9}
\]

We will use the characteristic function

\[
\varphi_X = E\left[e^{i\alpha x}\right] \tag{10}
\]

and the fact that the characteristic function of the sum of two random variables is the product of their characteristic functions. Using (10), the characteristic function of \(X\) is

\[
\varphi_X = E\left[e^{i\alpha x}\right] = \int_{-\infty}^{\infty} p_1 e^{i\alpha x} dx = \int_{-\infty}^{\infty} \sum_{j} \omega_j \frac{1}{\sigma_j \sqrt{2\pi}} e^{-\frac{(z-\mu_j)^2}{2\sigma_j^2}} e^{i\alpha x} dx \tag{11}
\]
The characteristic function of a Gaussian component with the mean $\mu$ and variance $\sigma^2$ is $\exp(i\mu + \sigma^2 t^2 / 2)$, so (11) can be simplified to

$$\phi_x = \sum_{j} \omega_j e^{i\mu_j - \frac{\sigma^2_{j}^2}{2}},$$

and the characteristic function of $Y$ can be represented as

$$\phi_y = \sum_{k} \omega_k e^{i\mu_k - \frac{\sigma^2_{k}^2}{2}}.$$  (13)

The characteristic function of $X+Y$ is equal to the product of the characteristic functions of $X$ and $Y$:

$$\phi_{x+y} = \phi_x\phi_y = \sum_{j} \sum_{k} \omega_j \cdot \omega_k e^{i[\mu_j + \mu_k] - \frac{\sigma^2_{j}^2 + \sigma^2_{k}^2}{2}},$$

which implies that the PDF of $Z$ is

$$Z : p_z = \sum_{j} \sum_{k} \omega_j \cdot \omega_k \frac{1}{\sqrt{(\sigma^2_{j} + \sigma^2_{k})2\pi}} e^{-\frac{[x - (\mu_j + \mu_k)]^2}{2(\sigma^2_{j} + \sigma^2_{k})}}$$

Equation (15) means that the density function of $Z$ is represented using Gaussian pairs from the GMMs of $X$ and $Y$. Each Gaussian component of $X$ forms a pair with a Gaussian component of $Y$. Each pair of Gaussian components generates a new Gaussian component of $Z$. The mean and variance of the newly-generated Gaussian component are respectively the sum of mean and variance of the Gaussians the pair. The weight of the new Gaussian component is the product of the weights of the Gaussians of the pairs. The new Gaussians of all pairs from $X$ and $Y$ form the GMM for $Z$. By exploiting these characteristics, the addition can be simply implemented in SLA-GMM without any approximation. This advantage slows the accumulation of error during the summation step.

IV. EXPERIMENTAL RESULTS

A. EXPERIMENTAL ENVIRONMENT

The proposed leakage analysis method was compared with existing methods. As a benchmark method, we used the most recent sampling-based method [12] which is the most effective smart sampling [11]. We also compared the state-of-the-art analytic model-based method, which uses the generalized extreme value distribution [18]. One hundred thousand MC data were used as the basis for evaluation of the accuracy of SLA-GMM and the benchmark methods.

In the following experiments, “16-nm predictive technology model high-performance” was used as the transistor model [29]. Variations in gate length, width, oxide thickness, and threshold voltage were considered. The $3\sigma$ values of the process parameter distributions were set to 18% of their mean values. For all benchmark methods, the BSIM4 model [20] was used as the leakage model, and HSPICE [30] was used as the leakage simulator. This is to exclude the impact on the accuracy of the leakage model when comparing the accuracy of each method. Ten ISCAS 85 circuits were used as the benchmark circuits.

Gaussian mixture modeling using the EM algorithm was implemented in Python language, and was based on open sources related to the machine learning algorithms provided by Tensorflow [31]. The other processes of SLA-GMM and the benchmark methods were implemented in C++. All implementations were done on Intel(R) Xeon(R) CPU E5-2690 @ 2.90 GHz.

B. VALIDATION OF THE ESTIMATION OF THE LEAKAGE DISTRIBUTION USING GMM

1) ACCURACY COMPARISON

We used goodness-of-fit statistics to quantify the accuracy of SLA-GMM, and verified the WB index by comparing the estimated $\kappa$ with the actual optimal number of clusters.

Firstly, the PDFs of the golden MC simulation, Quasi-MC (QMC), the GEV distribution, and SLA-GMM were graphically compared (Fig. 7a-c). The required number of QMC data is not defined, so the same number of data used for SLA-GMM was used for the QMC simulation and for fitting to the GEV distribution. The bin size of the QMC histogram was very large as expected because the bin size is proportional to the number of data. Therefore, even though the QMC helps improve the convergence rate of the leakage data, the QMC did not accurately estimate the PDF. Also, the GEV distribution was far from the actual leakage distribution, even though the GEV distribution has been known as the most accurate continuous function to approximate the leakage distribution. This failure occurred because the actual leakage distribution changes as the technology or supply voltage varies. Therefore, a fixed continuous function is not appropriate to estimate the leakage distribution.

The PDF of SLA-GMM consists of several Gaussian components (Fig. 7d-f). Thus, SLA-GMM can represent any functions, unlike a method that uses on class of continuous function. Therefore, even if the technology or the operating conditions such as supply voltage change, SLA-GMM can estimate the leakage distribution adaptively.

The shape of leakage distribution changes as the supply voltage changes (Fig. 8). However, the existing methods based on analytic models assume that the leakage distribution follows a certain fixed distribution function, and therefore they cannot estimate leakage distribution whose shape changes; in contrast SLA-GMM is resilient to changes of the shape of leakage distribution because the GMM can represent any continuous function. Under high or low supply voltage, SLA-GMM estimated the leakage distribution accurately (Fig. 8).
The Chi-square statistic $\chi^2$ and the Kolmogorov-Smirnov (K-S) statistic $D$ were used to quantify the similarity between two distributions by comparing the PDF and cumulative density function (CDF), respectively [10] (Table I). Small values of $\chi^2$ and K-S $D$ indicate accurately-estimated distributions. Average $\chi^2$ of SLA-GMM was $1.28 \times 10^{10}$, which is 98.3% and 98.6% smaller than those of GEV and QMC, respectively. Similarly, the average K-S $D$ of SLA-GMM was 0.01, which is 94.0% smaller than that of GEV, and comparable to that of QMC. These results showed that SLA-GMM is the most accurate of these methods to estimate the PDF and the CDF. K-S $D$ of our method is expected to be much smaller than that of the QMC after the summation operation, because $\chi^2$ has a large influence on the accuracy of the summation step. This effect will be described in the next section.

We chose the WB index as the cluster validation index. In our experiments, the trend of the WB index followed the trend of goodness-of-fit test results (Fig. 9). Small WB index, $\chi^2$, and K-S $D$ mean good estimation, so we can determine the number of clusters that reduce the goodness-of-fit statistics by using the smallest WB index value; i.e., we first use
the WB index to select the number of clusters (filled circle, Fig. 9). This process yields clustering results that have goodness-of-fit that are not far from the real solution (filled diamond, Fig. 9).

We tabulated (Table II) the numerical evaluation results of Fig. 9 for all benchmark circuits. The optimal number of clusters determined by the WB index increased the statistics by < 8%. Therefore, the WB index is effective to determine the optimal number of clusters when goodness-of-fit results are unknown.

2) Runtime

Runtime of SLA-GMM consists of data-collection time and GMM-modeling time. The number of QMC data was set as the same as that determined by SLA-GMM, so data-collection time required by the QMC was equal. The GMM modeling time was divided into clustering, cluster validation, and time required to calculate the required number of data. Data clustering was performed repeatedly until optimal numbers of clusters and data were determined. Generally, the one-time clustering time was several seconds, and increased as the number of clusters and data increased.
The accurate results of SLA-GMM were compared to the data-collecting time. This overhead of the achieved at the expense of 19.27% average runtime overhead GMM modeling decreased as the circuit size was increased, because the data-collecting time increases linearly as the circuit size increases, whereas the modeling time does not increase. Considering the increased accuracy of SLA-GMM, this overhead is less than the cost of collecting additional data.

### 3) VALIDATION OF SEQUENTIAL ADDITION

SLA-GMM is effective to analyze the leakage distribution of a system that consists of several sub-blocks. To show the effectiveness of the sequential addition, we used a partitioning algorithm to break ISCAS 85 circuits into several virtual blocks.

![Table II](image)

**Table II**

<table>
<thead>
<tr>
<th>Circuit</th>
<th>Chi-square statistic</th>
<th>K-S statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td># of clusters (A)</td>
<td>Relative $\chi^2$ at A</td>
</tr>
<tr>
<td>c17</td>
<td>9</td>
<td>0.0456</td>
</tr>
<tr>
<td>c432</td>
<td>9</td>
<td>0.0352</td>
</tr>
<tr>
<td>c499</td>
<td>7</td>
<td>0.0257</td>
</tr>
<tr>
<td>c880</td>
<td>12</td>
<td>0.0317</td>
</tr>
<tr>
<td>c1355</td>
<td>10</td>
<td>0.0498</td>
</tr>
<tr>
<td>c1908</td>
<td>9</td>
<td>0.0344</td>
</tr>
<tr>
<td>c2670</td>
<td>9</td>
<td>0.0266</td>
</tr>
<tr>
<td>c3540</td>
<td>9</td>
<td>0.0343</td>
</tr>
<tr>
<td>c5315</td>
<td>7</td>
<td>0.0346</td>
</tr>
<tr>
<td>c6288</td>
<td>8</td>
<td>0.0775</td>
</tr>
<tr>
<td>c7552</td>
<td>9</td>
<td>0.0494</td>
</tr>
</tbody>
</table>

Data-collecting time was greater than the GMM-modeling time (Table III). The accurate results of SLA-GMM were achieved at the expense of 19.27% average runtime overhead compared to the data-collecting time. This overhead of the GMM modeling decreased as the circuit size was increased, because the data-collecting time increases linearly as the circuit size increases, whereas the modeling time does not increase. Considering the increased accuracy of SLA-GMM (Table I), this overhead is less than the cost of collecting additional data.

We used the MLPart for the multi-level min-cut partitioning [32]. We used the Fiduccia-Mattheyses algorithm to perform the top-level partitioning, and the heavy-edge matching method for coarsening. We use V-cycling, which is composed of the repeated clustering-partitioning-refinement procedure, to use a solution that had been generated by a previous execution. We used a clustering ratio of 1.3, which has been shown empirically to be optimal [33].

We used the three largest ISCAS 85 circuits in this experiment. We first used the idea in Section III-A to estimate the leakage distributions of sub-blocks. Then we sequentially added the leakage distributions of the sub-blocks only using the parameters of the GMMs (Section III-B) to obtain Gaussian components of the GMMs of two sub-blocks and the resulting final leakage distribution of the total system (Fig. 10). The summation of the separately-modeled GMM distributions accurately estimated the full-chip leakage distribution. The CDF of SLA-GMM was much closer to the MC simulation result than was CDF of the QMC (Fig. 11).
The summation of the distributions is performed by the convolution of PDFs, so the error of estimating the PDF of a sub-block directly affects the summation of the PDF. For this reason, the QMC, which showed higher $\chi^2$ than SLA-GMM (Table I) showed abrupt increase in $D$ (> 0.6) after just one summation, compared to $D$ of SLA-DMM (Table IV). In addition, as the number of summation operations increases, the error of the estimate of the PDF increases so the accuracy of the summation generally degrades. Due to the low error of SLA-GMM in each summation step, SLA-GMM slowed the rate of degradation of the accuracy compared to QMC, when leakage distributions of two or more sub-blocks were summed (Table IV).

We compared the runtime of the QMC and SLA-GMM including the sequential addition module (Table IV). We used the most common fast convolution algorithm which uses the fast Fourier transform. Also, we parallized the process to use the same number of cores as the number of sub-blocks. Although the runtime for the summation of the QMC was less than that of SLA-GMM, the summation of the QMC must be performed every time the data points of interest change. However, the summation results are represented in an analytic function of GMM in SLA-GMM, so it does not require additional operations even when the data points change. In addition, the results showed that SLA-GMM estimated the summation results with lower K-S $D$ than QMC, which imposing only 2.64 % overhead on average, compared to QMC (Table IV). These results show that SLA-GMM has the high scalability and therefore can be used in SLA on a large circuit that consists of sub-blocks.

V. CONCLUSION

We proposed a method to use the Gaussian mixture model (GMM) in statistical leakage analysis (SLA). SLA-GMM consists of various shapes of leakage distribution modeling and sequential addition. The GMM can represent arbitrary functions and can be added sequentially with no assumptions,
TABLE IV
RESULTS OF K-S STATISTIC AND RUNTIME PROFILING OF THE SEQUENTIAL ADDITION

<table>
<thead>
<tr>
<th>Circuit</th>
<th># of sub-blocks</th>
<th>QMC</th>
<th>Proposed</th>
<th>Overhead (B-A)/A [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>C5315</td>
<td>0</td>
<td>0.012</td>
<td>11,052</td>
<td>11,052</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.834</td>
<td>7,500</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.786</td>
<td>5,040</td>
<td>0.148</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>0.901</td>
<td>2,160</td>
<td>0.306</td>
</tr>
<tr>
<td>C6288</td>
<td>0</td>
<td>0.014</td>
<td>35,388</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.688</td>
<td>8598</td>
<td>0.0603</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.813</td>
<td>5523</td>
<td>0.171</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>0.992</td>
<td>2758</td>
<td>1.069</td>
</tr>
<tr>
<td>C7552</td>
<td>0</td>
<td>0.013</td>
<td>14,220</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.839</td>
<td>11,820</td>
<td>0.0404</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.893</td>
<td>6,660</td>
<td>0.105</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>0.899</td>
<td>2,640</td>
<td>0.196</td>
</tr>
</tbody>
</table>

so it was considered to be a suitable model of leakage distribution. We also proposed a method based on the error bound of Gaussian distribution to calculate the required number of data. In experiments, SLA-GMM reduced $\chi^2$ by 98% and K-S D by 94% compared to the GEV distribution. The advantage of accuracy was achieved with 19.27% runtime overhead on average, compared to generating leakage data. Experiments on the sequential addition module showed the scalability of SLA-GMM. SLA-GMM reduced the accuracy loss more than the QMC did. Using the incremental and parallel characteristics, SLA-GMM can for SLA analysis of large circuits that consist of sub-blocks.

REFERENCES


