Adaptive Segmentation Gaussian Mixtures Models for Approximating to Drastically Scaled-Various Sloped Long-Tail RTN Distributions

Worawit Somha and Hiroyuki Yamauchi

Abstract—This paper proposes a fitting method to approximate the mixtures of various sloped-tail Gamma distribution characterizing the random telegraph noises (RTN) by an adaptive segmentation Gaussian mixtures model (GMM). The concepts central to the proposed method are 1) adaptive segmentation of the long-heavy tailed distributions such that the log-likelihood of GMM in each partition is maximized and 2) copy and paste with an adequate weight into each partition. This allows the fitting model to apply various bounded tail distribution even with multiple convex and concave folding curves. It is verified that the proposed method can reduce the error of the fail-bit predictions by 2-orders of magnitude while reducing the iterations for EM step convergence to 1/16 at the interest point of the fail probability of $10^{-12}$ which corresponds to the design point to realize a 99.9% yield of 1Gbit chips.

Index Terms—Mixtures of Gaussian, random telegraph noise, em algorithm, heavy-tail distribution, long-tail distribution, fail-bit analysis, static random access memory, guard band design.

I. INTRODUCTION

The approximation-error of the tails of random telegraph noise (RTN) distribution will become an unprecedentedly crucial challenge resulting from the fact that: (1) its error directly leads to the error of the guard band (GB) design required to avoid the out of spec after shipped to the market, and (2) tails of RTN distribution will become much longer than that of random-dopant-fluctuation (RDF) which is the conventional dominant factor of the whole margin-variations and the convolution results of the two will be more affected by the RTN than the RDF, as can be seen in Fig. 1. Since the increasing paces of variation-amplitude $V_{th}$ are differently dependent on the MOSFET channel-size (LW) like the below expressions of (1) and (2), the $V_{th}$ increasing paces of RTN is a 1.4x faster than that of RDF if assuming the LW is scaled down to 0.5 every process generation, as shown in Fig. 1(a).

\[
\Delta V_{th} \propto AV_{t} (RDF) \sqrt{LW} \tag{1}
\]
\[
\Delta V_{th} \propto AV_{t} (RTN) / LW \tag{2}
\]

where $AV_{t} (RDF)$ and $AV_{t} (RTN)$ are Pelgrom coefficients for RDF and RTN, respectively.

This means that RTN will soon exceed RDF and becomes a dominant factor of whole margin variations, as shown in Fig. 1(a). According to the references [1]-[5], there will come the time soon around 15nm scaled CMOS era.

To make clear the issues we will discuss in this paper, the concepts of what will happen at that time are shown in Figs. 1-2. Fig. 1(b) illustrates the probability density functions for RDF, RTN1(<40nm), RTN2(<16nm) and RTN3 (<7nm), and its convolution results, respectively.

It is worth mentioning that the distribution-shape of the convolution results obey the Gaussian when RTN<RDF and changes to follow the combinations of Gamma and Gaussian distributions when RTN=RDF, and finally becomes dominated by Gamma distribution of RTN when RTN>RDF, respectively [4]-[5]. The tails on the both sides of the distribution are asymmetrical and are differently influenced by longer-tail Gamma-RTN for right side and shorter tail Gaussian-RDF for left side and, respectively, as shown in Fig. 1(b).

Since the interest area for the GB design is on the right side, i.e., in less margin zone, it can be seen that the approximation-error of the RTN distribution directly leads to estimation-error of fail-bit counts (FBC). The conventional
Gaussian-model characterizing for the whole-margin variation can’t be used any more for analyzing such non Gaussian long-tail distributions of RTN. Fig. 2 shows how the affects of the approximation-error on the FBC error will be increased as the process dimension is scaled down. Until 15nm, its impact can be increased by 6 orders of magnitude compared to that of 40nm, as shown in Fig. 2.

![Fig. 2. Increased impact of approximation-error on the trouble of excessive under-estimation/over-estimation of the fail-bit counts.](image)

In order to solve the above issues, we propose, for the first time, a fitting method to approximate a long-tailed RTN distribution by an adaptive segmentation Gaussian mixtures model (GMM). This provides the following benefits: 1) applicable to the various convex and concave shapes of bounded Gamma distribution even with the wide range of shape-parameter $\alpha =0.02$ to $1.15$ and 2) still using Gaussian distribution to simply utilize an error-function for cumulative density function. The main contribution of this paper is to point out that it is possible to approximate the long tailed distributions by mixtures of convenient Gaussian probability distributions, so that available yield-prediction models can be effectively analyzed and so that the effect of the long tailed distributions upon the fail-bit count accuracy can be analytically determined. This is because the convolution result of linear combinations of Gaussians becomes also analytically determined. This makes it easier to predict the fail-bit counts before and after screening at the stages of both circuit design and screening test [1]-[3].

Here is how the rest of this paper is organized. In Section II, we refer to some of the example as evidence indicating how the conventional models cause intolerable huge error to make clear the purpose of the proposed work. In Section III, we will propose our recursive algorithm for constructing approximating Gaussian mixtures model (GMM). In Section IV, we refer to some of the example as evidence indicating if the proposed models can approximate well the heavy long-tailed distributions. We give a precise fail-bit count prediction. In Section V, we rigorously prove that it is possible to approximate various long-tailed distributions with bounded convex and concave curves by mixtures of Gaussian distributions. Finally, we state our conclusion in Section VI.

II. DISCUSSIONS ON THE CONVENTIONAL MODELS

Expectation-maximization (EM) algorithm [6], which is an iterative procedure that maximizes the likelihood of Gaussian mixtures models (GMM), is well known as easy and convenient means to approximate GMM to the non Gaussian distributions.

III. PROPOSED STATISTICAL APPROXIMATION MODEL FOR RTN GAMMA DISTRIBUTION

In order to solve these crucial issues, we develop a remarkably simple adaptively segmentation EM algorithm-based fitting algorithm. The centerpiece of this idea is: (a) adaptive partitioning of the long tailed distributions such that the log-likelihood of GMM is maximized in each segmentation and (b) copy and paste fashion with an adequate weight into each partition for constructing the whole long-tail distributions. The concepts of the two different proposed EM-based approximation means are shown in Fig. 4(a) and (b), respectively.

A. Adaptive Segmentation

Algorithm of the adaptive segmentation is described below from step 1) to step 4).

1) $1^{st}$-step is to do approximation by 3-GMM between $X_0$ and $X_n$. And find the point of $X_1$, where likelihood of 3-GMM is maximized.
2) 2nd-step is to do the same thing as 1) between X1 and Xn. And find the point of X2, where likelihood of 3-GMM is maximized.

3) 3rd-step is to do the same thing as 2) between X2 and Xn. And find the point of X3, where likelihood of 3-GMM is maximized between X3 and Xn.

This flow can be repeated until the likelihood of whole GMM can be maximized as shown in Fig. 4(a).

B. Copy and paste fashion

Algorithm of the copy and paste fashion is described below from step 1) to step 4).

1) 1st-step is to do approximation by 3-GMM between X0 and Xn. And find the point of X1, where likelihood of 3-GMM is maximized. X is given by (X1-X0) and w0 is the weight of the 1st 3-GMM.

2) 2nd-step is to get the weight (w1) of the 2nd 3-GMM. And copy the 1st 3-GMM and paste it into the adjacent place (shifted by X) by weighting of \( \Delta w_1 \), which is given by the slope of Gamma distribution.

where, slope = \( (w_0 - w_1) / X \)

3) 3rd-step is to do the same thing as 2), as shown in Fig. 4(b). This flow can be repeated until Xm>Xn.

This algorithm can allow approximating any heavy-long tailed distributions by the convenient short-tail Gaussian probability distributions. Even if the whole distributions are comprised of mixtures of various convex and concave curves as shown in Fig. 4(c), individual area of (O-P), (P-Q), (Q-R), (R-S), and (S-T) can be adaptively segmented based on its slope. It is a clear that the both proposed ideas can apply to this kind of distribution. In Section V, an example of actual distributions of future RTN is given and discussed.

Thanks to the segmentation, the range of variables for the 3-GMM approximation is limited and almost similar to the other segmentations. This can make the number of EM-iterations required to find the best point smaller and help to avoid the wrong convergence point unlike the conventional EM-algorithm, as shown in Table I. This also allows us to use only Gaussian distributions when doing convolution of Gaussian-RDF and Non-Gaussian-RTN distributions.

<table>
<thead>
<tr>
<th>TABLE I: COMPARISONS OF EM-ITERATIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Segmentation (proposed)</td>
</tr>
<tr>
<td>Gamma1</td>
</tr>
<tr>
<td>Gamma2</td>
</tr>
<tr>
<td>Gamma3</td>
</tr>
</tbody>
</table>

The convolution results also can be given by analytical simple and convenient expressions of just linear combination of Gaussian, which can give us the fail-bit count by just summing up the values of normal (Gaussian) cumulative density function (normcdf) for each Gaussian of the whole GMM. The example of how to calculate the fail-bit error counts of the segmentation of \( (x_a-x_b) \) is shown in Fig. 5.
IV. DISCUSSION ON ACCURACY OF STATISTICAL APPROXIMATION MODEL FOR RTN DISTRIBUTION

To illustrate the effects of the proposed scheme on the reduction of the approximation-error in the long tails, the following examples are assumed: (1) ratio of how fast does the tail decay of Gaussian-RDF and Gamma-RTN, i.e., its parameters are assumed as follows: ($\alpha=1, \beta=0$) for Gaussian, ($\alpha=1, \beta=0.56$) for Gamma. The relationship between the two distributions and its convolution are shown in Fig. 1b) and (2) comparisons of the following 6 approximation-models of Gamma distribution ($\alpha=1, \beta=0.56$): (a) the number of 3, 9, 24, and 128 of GMMs are used for fitting the whole distribution (no segmentation) and (b) the number of 3 and 9 of GMMs are used for approximating each segmentation comprising whole distribution.

Fig. 6 shows the comparisons of the convolution results between the cases with and without segmentation schemes. It is found that 3-GMM segmentation scheme can reduce the errors by 10$^5$, 10$^6$, and 4x than 3, 9, and 24-GMMs, respectively. As mentioned earlier, 9-GMM segmentation is worth than 3-GMM segmentation in the wide range of $x=-6\sim-12$ because the variation of probability becomes larger and the density of GMM in area of lower probability becomes much smaller as shown in Fig. 7.

V. APPLICATION TO MORE COMPLEX DISTRIBUTIONS

According to the reference [7]-[9], the distribution of RTN amplitude will have a complex bounded tail caused by “atomistic” variation-behaviors with various variation factors of the gate line-edge roughness (GER), fin-edge roughness (FER), and metal gate granularity (MGG), as shown in Fig. 9. They are no longer obeyed to the single gamma distribution but to the multiple gamma distribution depending on the tail positions of (O-P), (P-Q), and (Q-R), as shown in Fig. 9. As the examples to illustrate the effectiveness of the proposed fitting models, the three types of distributions whose have a
different folding points are given as Combo1, Combo2, and Combo3, as shown in Fig. 9.

The proposed both ideas of “adaptive segmentation” and “copy and paste” fashion can apply to this kind of complex non-linear distribution. This is because the width of each segmentation is much smaller than the length of (O-P), (P-Q), and (Q-R). The same concepts can be used in each partition of (O-P), (P-Q), and (Q-R).

Fig. 10 shows the comparisons of approximation-errors for fitting to Combo1, Combo2, and Combo3 between the cases of (a) using the conventional 3-GMM model and (b) using the proposed segmentation models.

It is found that the trend of cdf errors depending on the margin scale of x position is similar between the different distributions of Combo1-3, as can be seen in Fig. 11. The cdf errors for the “copy and paste” are smaller than that for the “adaptive segmentation” in the smaller x-position. Contrary, its relationship is inverted. Since the region of a larger x and a smaller probability like $10^{-12}$ is more interest area for the rare event fail-bit count analyses, it can be said that the proposed idea of “adaptive segmentation” provides the better fitting model to predict the yield-loss after shipped to the market due to the time-dependent RTN-caused failures.

VI. CONCLUSION

In this paper, we have discussed, for the first time, how the various-sloped RTN distribution-tail should be approximated and how much its approximation-error can affect on the accuracy of the statistical predictions of the number of fail-bit counts, which is required to avoid the out of spec after shipped to the market. It has been pointed out that the conventional Gaussian models can’t be used any more due to intolerable model errors caused by the deviation from the actual RTN-caused distributions, once the distribution-tail of the RTN becomes longer than that of the conventional variations of the RDF. This is because the tail of convolution results doesn’t obey to the Gaussian any more but follows to the mixtures of various-sloped Gamma distributions.

To address the above issues, we have proposed the two types of an effective simple algorithm for approximating the tails of RTN distributions by convenient and simple GMM. This allows the fitting model to apply the various bonded tail distributions even with the multiple convex and concave folding curves. It has been verified that the proposed method can reduce the error of the fail-bit predictions by 2-orders of magnitude while reducing the iterations for EM step convergence to 1/16 at the interest point of the fail probability of $10^{-12}$ which corresponds to the design point to realize a 99.9% yield of 1Gbit chips.

We have also pointed out that the proposed scheme is a candidate fitting algorithm for the distributions of the future RTN distributions, which will be crucial not only for the circuit design but also the GB design for screening test when RTN variables becomes larger than that of RDF.
ACKNOWLEDGMENT

The authors are grateful to Yan Zhang, Yu Ma for their helps.

REFERENCES


Worawit Somha received master degree in electrical engineering from King Mongkut’s Institute of Technology Ladkrabang (KMITL), Bangkok, Thailand. His master thesis was on “Vector Quantizers for Speech Coding and there Implementation on TMS-320C30”. Since 1995 he has given a lecture for bachelor degree student in subject of “Introduction to Digital Signal Processing” at KMITL as an assistance professor, and his research area is speech coding. Since 1997 he has worked with the company in the position of consulting engineering. Since 2012 he has got the scholarship from KMITL for D.Eng. student program and now being pursuing PhD degree in major of intelligence information system engineering at Fukuoka Institute of Technology.

Hiroyuki Yamauchi received the Ph.D. degree in engineering from Kyushu University, Fukuoka, Japan, in 1997. His doctoral dissertation was on “Low Power Technologies for Battery-Operated Semiconductor Random Access Memories”. In 1985 he joined the Semiconductor Research Center, Panasonic, Osaka, Japan. From 1985 to 1987 he had worked on the research of the submicron MOS FET model-parameter extraction for the circuit simulation and the research of the sensitivity of the scaled sense amplifier for ultrahigh-density DRAM's which was presented at the 1989 Symposium on VLSI Circuits. From 1988 to 1994, he was engaged in research and development of 16-Mb CMOS DRAM's including the battery-operated high-speed 16 Mbit CMOS DRAM and the ultra low-power, three times longer, self-refresh DRAM which were presented at the 1993 and 1995 ISSCC, respectively. He also presented the charge-recycling bus architecture and low-voltage operated high-speed VLSI's, including 0.5V/100MHz-operated SRAM and Gate-Over-Driving CMOS architecture, which were presented at the Symposium on VLSI Circuits in 1994 and 1996, respectively, and at the 1997 ISSCC as well. After experienced general manager for development of various embedded memories, eSRAM, eDRAM, eFlash, eFeRAM, and eReRAM for system LSI in Panasonic, he has moved to Fukuoka Institute of Technology and become a professor since 2005. His current interests are focused on study for variation tolerant memory circuit designs for nano-meter era. He holds 212 Patents including 87 U.S. Patents and has presented over 70 journal papers and proceedings of international conferences including 10 for ISSCC and 11 for Symposium on VLSI Circuits. Dr. Yamauchi received the 1996 Remarkable Invention Award from Science and Technology Agency of Japanese government and the highest ISOCC2008 Best Paper Award.

He had been serving a program committee of ISSCC for long periods, from 2002 through 2009. He served a program committee of IEEE Symposium on VLSI Circuits from 1998 through 2000 and has come back and been serving again since 2008. He is also serving A-SSCC since 2008.