A Look up Table Based Adaptive Segmentation Gaussian Mixtures Model for Fitting Complex Long-Tail RTN Distributions

Worawit Somha and Hiroyuki Yamauchi

***Abstract*—This paper proposes a look up table (LUT) based fitting method to approximate the mixtures of various sloped-Gamma tail distributions by an adaptive segmentation Gaussian mixtures model (GMM). The concepts central to the proposed method are 1) LUT based fitting of all parameters of GMM and segmentation width and 2) adaptive segmentation of the long tailed distributions such that the log-likelihood of GMM in each partition is maximized. This allows the LUT based GMM model to apply any arbitrary shaped tail distribution even with multiple convex and concave folding curves, while eliminating the need of any EM iterations. It is verified that the LUT based GMM model can reduce the error of the fail-bit predictions by 2-orders of magnitude 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, look up table, 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)-caused distribution will become an unprecedentedly crucial challenge resulting from the facts that: (1) complex tails of RTN distributions itself will dominate the overall convolution results, as can be seen in Fig. 1, because its tails of the variations will become much longer than that of random-dopant-fluctuation (RDF) which is the conventional dominant factor of the whole margin-variations and (2) its errors directly lead to the error of the guard band (GB) design required to avoid the out of spec after shipped to the market.

According to the reference [1]-[5], the distribution of the 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. 1. They are no longer obeyed to the single gamma distribution but to the mixtures of different sloped-gamma distribution depending on the tail positions of (O-P), (P-Q), and (Q-R), as shown in Fig. 1. Since the increasing paces of variation-amplitude ΔVth of the RTN can be a 1.4x faster than that of the RDF based on the below expressions of (1) and (2), the RTN will soon exceed the RDF and becomes a dominant factor of the whole margin variations, as shown in Fig. 1. According to the references [1]-[5], there will come the time soon around 15nm scaled CMOS era.

$$\Delta\text{Vth (RDF)} \propto \text{AVt (RDF)} \Big/ \sqrt{\text{LW}} \qquad (1)$$

$$\Delta\text{Vth (RTN)} \propto \text{AVt (RTN)} \ / \ \text{LW} \qquad (2)$$

where LW is the MOSFET channel-size (LW) and AVt(RDF) and AVt(RTN) are Pelgrom coefficients for RDF and RTN, respectively. The ratio of the increasing pace of RTN to RDF is given by SQRT(0.5)/0.5=1.4 assuming that scaling pace of LW is 0.5x per generation.

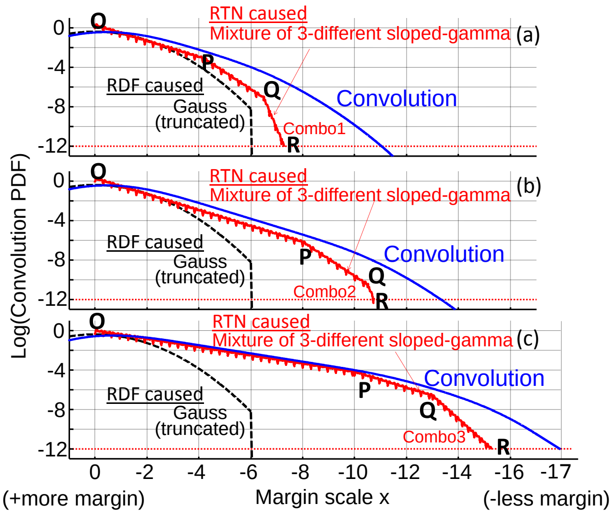


Fig. 1. Comparisons of convolution results between the truncated Gaussian RDF and 3-different gamma RTN distributions: (a)combo1,(b)combo2, and (c) combo3.

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 illustrates the probability density functions (pdf) for truncated RDF, 3-different complex distributions of the RTN amplitude, and its convolution results, respectively.

Since the pdf of the rare event zone (pdf < 10^{-12}) is almost governed by the RTN distribution, its approximation errors of the RTN distribution directly lead to an estimation error of the fail-bit counts (FBC). The conventional Gaussian model [6]-[8] characterizing for the whole-margin variation can't be used any more for analyzing such complex mixture of Gamma long-tail distributions of the RTN.

Since the overall shape of RTN distributions can be modulated by the contribution-ratio of the "atomistic" variation-behaviors of GER, FER, and MGG [1], we have to assume any arbitrary distributions whose have multiple

Manuscript received December 25, 2012; revised March 8, 2013. This work was supported in part by MEXT/JSPS KAKENHI Grant Number of 23560424 and grant from Information Sceience Laboratory of Fukuoka Institute of Technology.

The authors are with the Information Intelligent System Fukuoka Institute of Technology, 3-30-1, Wajiro-Higashi, Higashi-ku, Fukuoka, Japan (e-mail: bd12002@ bene.fit.ac.jp, yamauchi@fit.ac.jp).

DOI: 10.7763/IJMMM.2013.V1.52

convex and concave folding points (O-Z), as shown in Fig. 2(a). Each segmentation (O-P, P-Q, --, Y-Z) can be characterized by the slope parameter of the gamma distribution of β1-βn, as shown in Fig. 2(b). However, the appropriate approximation method for meeting the requirements above have not been proposed yet.

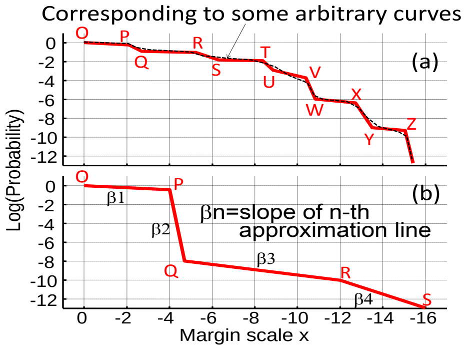


Fig. 2. (a) Mixtures of multiple convex and concave folding points, which can approximate any arbitrary shaped curves. (b) long tail distributions of mixture of different sloped gamma distributions of β1(O-P),β2(P=Q),β3(Q-R), and β4(R-S).

In order to solve the above issues, we propose, for the first time, a look up table (LUT) based fitting method to approximate any arbitrary 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 β=0.05 to 0.95 while eliminating the need of EM iterations and (2) still using Gaussian distribution to simply utilize a normal cumulative density function for calculating the FBC. The main contribution of this paper is to point out that it is possible to approximate any shaped long tailed distributions by the LUT-based fitting 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 FBC accuracy can be analytically determined. This is because the convolution result of linear combinations of Gaussians becomes also Gaussians and can be expressed by the analytical expressions, which allows using normal (Gaussian) cumulative density function (normcdf) for estimating the error counts. This makes it easier to predict the FBC before and after screening at the stages of both circuit design and screening test.

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 the LUT-based fitting method with our recursive algorithm for constructing approximating GMM. In Section IV, we refer to some of the example as evidence indicating if the proposed models can approximate well any mixtures of various sloped tail distributions. We give a precise FBC prediction. In Section V, we rigorously prove that it is possible to approximate the various long-tailed distributions with the 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 [9], 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.

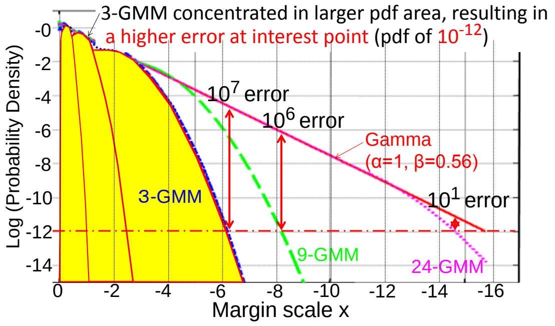


Fig. 3. Increased impact of approximation-error on the trouble of excessive under-estimation/over- estimation of the fail-bit counts.

However, all GMMs given by this fitting algorithm tend to concentrate in the non-tail region in which the sensitivity to increase the likelihood is much larger than that for the tail region, as shown in Fig. 3. Since the interest region for analyzing the FBC of the rare-events is in the tail region (at probability of 10^{-12}), the EM algorithm for this application leads to a significant FBC error of orders of 10^{7}, as shown in Fig. 3. Even if increasing the number of GMM from 3 to 9 and 24, the significant error of orders of 10^{6} and 10^{1}, respectively, are still remained, as shown in Fig. 3. In almost all FBC analyses, the distribution of interest only matters in the tail-region of probability of orders of 10^{-12} [6]-[8]. Thus, this is a crucial challenge we should address until the time comes for the rare-event SRAM yield predictions.

III. Proposed LUT Based Statistical Approximation Model for RTN Gamma Distribution

In order to solve these crucial issues, we develop a remarkably simple LUT-based 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, (b) copy and paste fashion with an adequate weight into each partition for constructing the whole long-tail distributions and (c) all the parameters required to regenerate GMM in individual segmentation are given by the pre-defined LUT for eliminating the need of any EM iterations. The concepts of the three different proposed EM-based approximation means are shown in Figs. 4(a)-(b) and Fig. 5, respectively.

A. *Adaptive Segmentation Based Fitting*

Algorithm of the adaptive segmentation 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.
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) 3^{rd}-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).

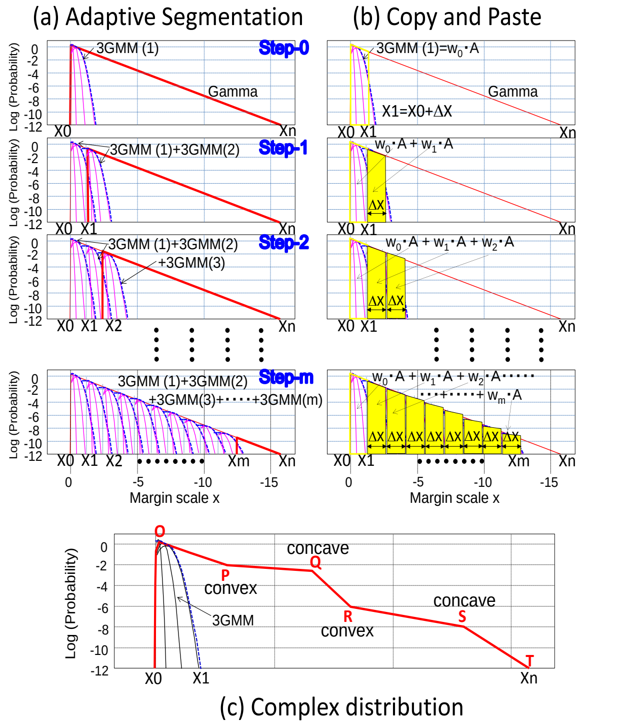


Fig. 4. Concepts of the proposed approximation algorithm. (a) adaptive segmenttaion:Xm are decided such that likelifood of each 3-GMM can be maximized. (b) copy and paste fashion: copy the first 3-GMM and paste to others with adaptive weighting. (c) example of complex distributions comprising various variation factors

B. *Copy and Paste Fashion Based Fitting*

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

1) 1^{st}-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. ΔX is given by (X1-X0) and w_0 is the weight of the 1^{st} 3-GMM.

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

where slope=(w_0 - w_1)/ ΔX

3) 3^{rd}-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 angled slope 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.

C. *Look up Table (LUT) Based Fitting*

However, as the number of folding points in increased, the number of EM operations required to get the GMM for the individual segmentation is increased. Thus, this paper also proposes the LUT-based GMM generating means to make this idea really practical by eliminating the need of EM iterations. This can eliminate any steps of EM operations. If the information of the slope of the individual segmentation, e.g., β of shaped parameter of gamma distribution is just input, the LUT outputs the all parameters required to regenerate the GMM comprising 3-Gaussians, as shown in Fig. 5. This also outputs the best width of individual segmentation ΔX(seg) that can the likelihood can be maximized. As a result, overall approximations with optimized segmentation width can be easily done without any time-consuming EM steps. In this paper, we assumed the range of slope is β=0.05～0.95, which corresponds to the variations of the slope of log-scaled gamma distributions, as shown in Fig. 6.

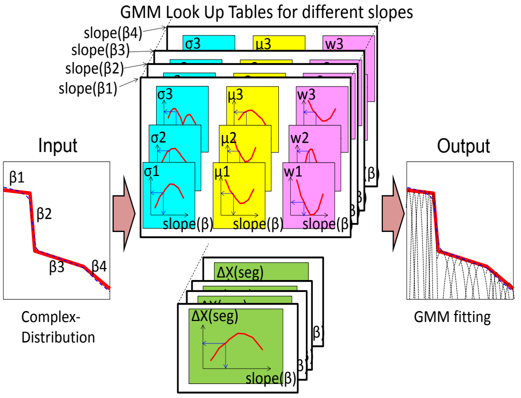
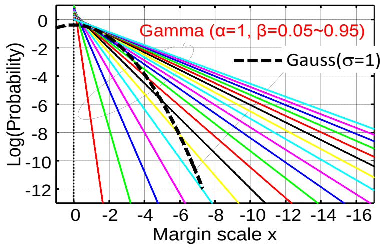


Fig. 5. Concept of the look up table for the different sloped- gamma distributions of shape parameter β. This LUT provides the parameter set of 3-GMMs (α,β,w) and width of segmentation which maximizes its likelihood in the segmentation.

Since the slope β dependencies of all the parameters of GMM and the best width of individual segmentation ΔX(seg) have a simple and continuous relationship, as shown in Figs. 7 and 8, the error cause by interpolation of LUT can be minimized as discussed in the following Section IV.



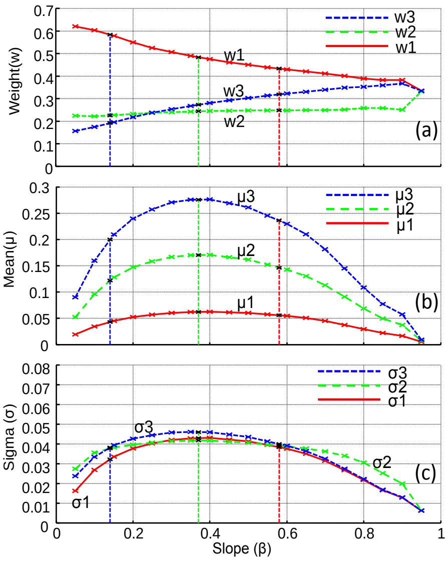


Fig. 7. Slope dependency of parameters in look up table for three Gaussian mixture model.

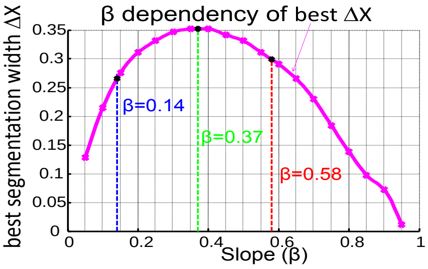


Fig. 8. Slope dependency of parameters in look up table for the best segmentation width ΔX.

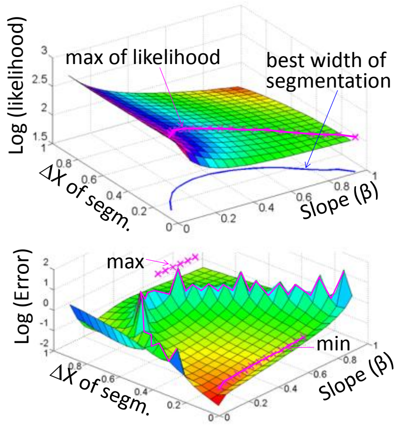


Fig. 9. Likelihood and approximation error dependencies of the slope-β and the segmentation width-ΔX.

Fig. 9 shows the positions of the maximum likelihood and the minimum of the approximation error in individual segmentation. The point of the best segmentation width ΔX depends on the slope β and corresponding to the point of the maximum likelihood, as shown in Fig. 9. Thus, if the slope β is input to the LUT, the ΔX is also given besides the parameter set for 3-GMMs (shown in Fig. 7).

IV. Discussion on Accuracy of Statistical Approximation Model for RTN Distribution

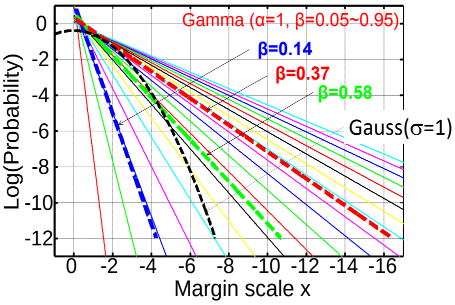


Fig. 10. 3-sampling points of slope β=0.14, 0.37, and 0.58.

To illustrate the effects of the proposed LUT based scheme on the approximation-error in the interest region, the following 3-examples of the different sloped gamma distribution are assumed: (α=1, β=0.14), (α=1, β=0.37), and (α=1, β=0.58), respectively. The relationships between the three different sloped-gamma and Gauss distributions are shown in Fig. 10.

Fig. 11 shows the 3-GMMs in the different segmentations of ΔX for the 3-different sloped tails of (a) ΔX=0.3, β=0.58, (b) ΔX=0.35, β=0.37, (c) ΔX=0.27, β=0.14, respectively. The LUT provides this kind of parameter set for regenerating 3-GMM and the best segmentation width ΔX.

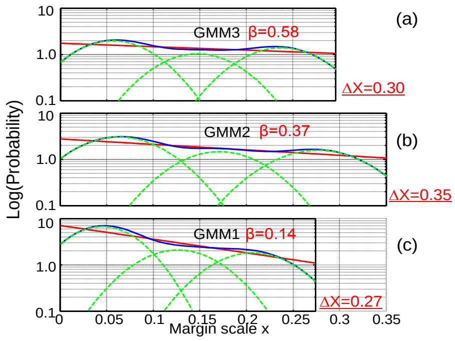


Fig. 11. 3-GMMs in the different best ΔX-segmentation for different sloped tails of (a)ΔX=0.3,β=0.58, (b) ΔX=0.35, β=0.37, (c) ΔX=0.27,β=0.14.

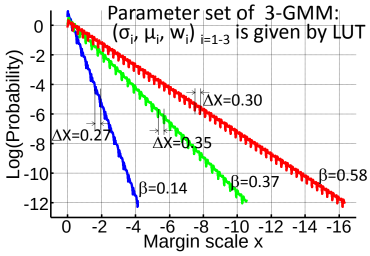


Fig. 12. LUT based fitting of the different sloped tails of (a)ΔX=0.3,β=0.58, (b) ΔX=0.35, β=0.37, (c) ΔX=0.27,β=0.14.

Fig. 12 shows that LUT-based fitting curves for the 3-different sloped gamma distributions of β=0.14, 0.37, and 0.58, respectively. The weight of the individual segmentation at each X-point is also given by the LUT.

To illustrate the effects of the proposed LUT based scheme on the approximation-error in the interest region, the errors of the cumulative density function (cdf) of the convolution results are compared between the proposed 3-schemes and the conventional one without any segmentation manners. Here, the convolutions are done between the 3-different sloped gamma distributions and Gauss distribution (σ=1), which are assumed the amplitude ratio relationship between the RTN and RDF variations [1]-[2].

Fig. 13 shows the cdf-error comparison results between the proposed 3-schemes and the conventional one without any segmentation manners. To make clear the effects of the proposed LUT based scheme on the approximation-error in the interest region compare with the other two proposed schemes, the orders of error are compared in the interest region (cdf of 10^{-12}), as shown in Fig. 13. It can be seen that the LUT can reduce the errors by the two orders of magnitude compared with the conventional schemes as well as the other two proposed schemes.

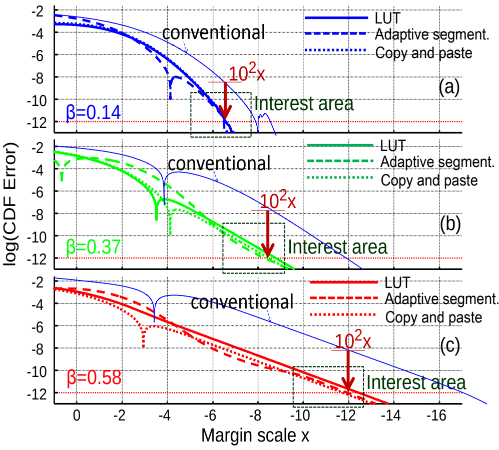


Fig. 13. Cdf error of the convolution results between Gauss (σ=1) and 3-different Gammas of (a) β=0.14, (b) β=0.37, and (c) β=0.58, respectively.

V. Application to More Complex Distributions

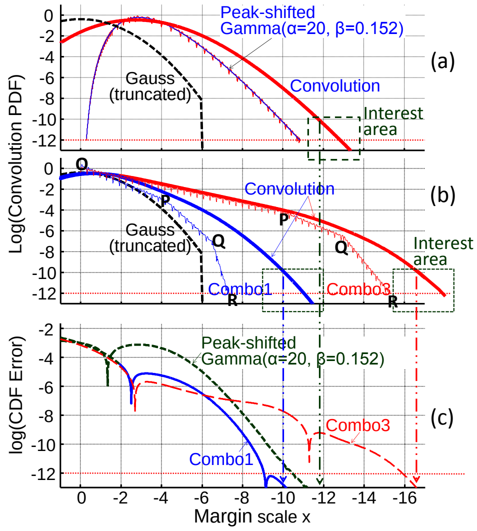


Fig. 14. Convolutions of (a) peak-shifted gamma and (b) 3-different sloped combined gamma distributions and the truncated Gauss distribution. (c) Cdf-error comparisons between the three different distributions of (a) and (b).

According to the reference [1], the distributions of RTN amplitude are no longer obeyed to a 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 Figs. 1 and 14. As the examples to illustrate the effectiveness of the proposed fitting models, the two types of distributions whose have a different folding points are given as Combo1 and Combo3, as shown in Fig. 14(b). In addition, the more complex distribution, whose peak position is shifted and tail distribution is deviated from the simple exponential functions, is also tried because the reference [1] uses such kind of shapes as an example of the potential future RTN distribution, as shown in Fig. 14(a).

It is verified that the proposed LUT based fitting can apply to any arbitrary sloped distributions even it has a complex and non linear distribution as Figs. 14(a) and (b) shows, while reducing the error of cdf to less than 1%, as shown in Fig. 14(c). As can be seen in Fig. 14, cdf-errors for the different three complex distributions are smaller than 10^{-12} at the point where pdf of the convolution results is 10^{-10}. It means that the error of the fail-bit count (FBC) is smaller than 1% at this kind of rare-event level. 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 LUT-based fitting scheme provides the practical fitting model to predict the yield-loss after shipped to the market due to the time-dependent RTN-caused failures. This can adapt any arbitrary sloped distributions without any need of computing power for the EM convergence unlike the two other proposed schemes.

VI. Conclusion

In this paper, we have proposed, for the first time, the LUT based GMM fitting schemes for approximating the complex combination of the various-sloped gamma distributions with multiple convex and concave folding points and discussed how much its approximation-error can affect on the accuracy of the statistical predictions of the number of the fail-bit counts, which is required to avoid the out of spec after shipped to the market. It has been pointed out that proposed LUT based fitting method can provide the practical fitting model to predict the yield-loss after shipped to the market due to the time-dependent RTN-caused failures. This can adapt any arbitrary sloped distributions without any need of computing power for the EM convergence.

We have proposed and compared the three types of an effective simple algorithm for approximating the tails of RTN distributions by convenient and simple GMM. It has been verified that the proposed LUT based method can reduce the error of the fail-bit predictions by 2-orders of magnitude at the interest point of the fail probability of 10^{-12} as well as the other two proposed schemes. The LUT based schemes can eliminate the need of any computing power for the EM iterations. This is the advantage over the two other proposed schemes.

Acknowledgment

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

References

[1] X. Wang *et al.*, "RTS amplitude distribution in 20nm SOI FinFETs subject to statistical variability," *SISPAD* 2012, pp. 296-299.

[2] X. Wang *et al.*, "Simulation Study of Dominant Statistical Variability Sources in 32-nm High-k/Metal Gate CMOS," *IEEE Electron Device Letters - IEEE ELECTRON DEV LETT*, vol. 33, no. 5, pp. 643-645, 2012.

[3] K. P. Cheung *et al.*, "The amplitude of random telegraph noise: Scaling implications," in *Proc. 2012 IEEE International Reliability Physics Symposium (IRPS)*, pp. 1.1 - 1.3.

[4] K. Takeuchi *et al.*, "Comprehensive SRAM Design Methodology for RTN Reliability," *Digest of IEEE Symposium on VLSI Technology*, 2011, pp. 130-131.

[5] K. Takeuchi *et al.*, "Direct Observation of RTN-induced SRAM Failure by Accelerated Testing and Its Application to Product Reliability Assessment," *Digest of IEEE Symposium on VLSI Technology*, 2010, pp. 189-190.

[6] H. Yamauchi, "A Discussion on SRAM Circuit Design Trend in Deeper Nanometer-Scale Technologies," *IEEE Transactions Very Large Scale Integration (VLSI) Systems*, vol. 18, issue 5, 2010, pp. 763-774.

[7] H. Yamauchi, "Embedded SRAM trend in nano-scale CMOS," *IEEE International Workshop on Memory Technology, Design and Testing, MTDT 2007*, pp. 19 – 22, 2007.

[8] H. Yamauchi and Tutorials "Variation tolerant SRAM circuit design," *IEEE ISSCC 2009 and IEEE A-SSCC 2008*.

[9] T. K. Moon, "The expectation-maximization algorithm," *IEEE Signal Processing Magazine*, vol. 13, issue 6, pp. 47 – 60, 1996.

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.