International Journal of Electrical, Electronic and Communication Sciences
ISSN: 2517-9438
Vol:7, No:3, 2013

A Nano-Scaled SRAM Guard Band Design with Gaussian Mixtures Model of Complex Long Tail RTN Distributions

Worawit Somha, and Hiroyuki Yamauchi

***Abstract*—**This paper proposes, for the first time, how the challenges facing the guard-band designs including the margin assist-circuits scheme for the screening-test in the coming process generations should be addressed. The increased screening error impacts are discussed based on the proposed statistical analysis models. It has been shown that the yield-loss caused by the misjudgment on the screening test would become 5-orders of magnitude larger than that for the conventional one when the amplitude of random telegraph noise (RTN) caused variations approaches to that of random dopant fluctuation. Three fitting methods to approximate the RTN caused complex Gamma mixtures distributions by the simple Gaussian mixtures model (GMM) are proposed and compared. It has been verified that the proposed methods can reduce the error of the fail-bit predictions by 4-orders of magnitude.

***Keywords*—**Mixtures of Gaussian, Random telegraph noise, EM algorithm, Long-tail distribution, Fail-bit analysis, Static random access memory, Guard band design.

I. Introduction

THE *guard band* (GB) designs including the *margin assist circuits* (MRASST) scheme for the *static random access memory* (SRAM) [1]-[3] will face an unprecedentedly crucial challenge in the coming process generations. This stems from the facts originated with that the *time-dependent* (TD) *margin-variations* (MV) after the screening will become much larger than that of ordinary *non*-TD-MV [4]-[8]. This trend indicates that the number of failures caused by the TD-MV after the screening will have a dominant influence over the whole yield loss unless adequately treated at the GB designs including the MRASST designs in the coming process generations. These failures can't be screened out by the ordinary functional test based on the conventional GB designs any more without a huge chip yield-loss. This results from the facts that it is really hard to predict the amount of the margin degradation of the SRAM operating voltage (Vdd) caused by the TD-MV during the guaranteed lifetime period.

The main reason behind the challenges facing the statistical predictions of the TD-MV caused failures is a big change of the statistical distribution of the whole MV from the simple Gaussian to the complex Gamma mixtures distributions. Since the *threshold voltage* (Vth) distribution is the dominant contributor to the MV, the trend of the Vth is explained in the following.

The Vth distributions of the nano-scaled CMOS have clearly shown that we have to consider not only the *non*-TD spatial *random dopant fluctuation* (RDF) but also the TD temporal Vth variations due to the *random telegraph noise* (RTN) [4]-[8]. It has been well shown in [4]-[8] that distributions for the *amplitude of Vth modulation* (ΔVth) due to RDF and RTN are obeyed to a Gaussian and a complex sloped Gamma mixtures distribution, respectively. In addition, the increasing paces of ΔVth amplitude are differently dependent on the MOSFET channel-size (LW) like the following (1) and (2).

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

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

where AVt(RDF) and AVt(RTN) are Pelgrom coefficients for RDF and RTN, respectively. Assuming the LW is scaled down to 0.5 every process generation, the ΔVth increasing paces of the RTN is a 1.4x faster than that of RDF. This means that the TD-ΔVth(RTN) will soon exceed the *non*-TD ΔVth(RDF) and becomes a dominant factor of the whole margin variations. According to [5]-[7], there will come the time soon around a 15nm scaled CMOS era.

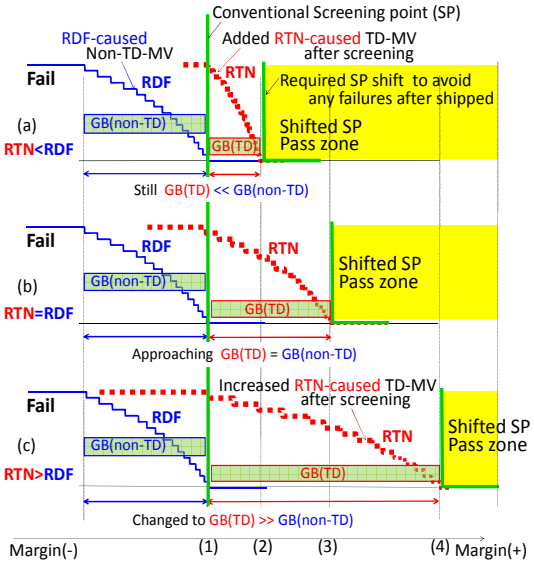

Fig. 1 Change of the GB width ratio for *non*-TD to TD variations caused by RDF and RTN: (a) RTN<RDF, (b) RTN=RDF, (c) RTN>RDF. Increased GB causes to increase the number of discarding chip at screening test

Manuscript received February 3, 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.

Worawit Somha[1] and Hiroyuki Yamauchi[2] 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[1] and yamauchi@fit.ac.jp[2])

This indicates that the GB designs including the MRASST designs should rely almost entirely on the statistical predictions of the amounts of TD-MV.

Because the required GB voltages will be no longer small fraction of the whole margins, as shown in Fig. 1, the conventional GB design criteria with the screening test won't be effective any more for avoiding the out of spec after the screening.

To make clear the issues we will address in this paper, the concepts of what will happen in the coming process generations are shown in Figs. 1 and 2. The GB(TD) in Fig. 1 refers to the GB voltages corresponding to the shifting amount of TD-MV due to RTN. The required GB(TD) for avoiding the out of spec becomes larger due to ever increased RTN and exceeds soon the GB(*non*-TD) for RDF. As a result, the number of discarding chips after the screening will be no longer neglected, as shown in Figs. 1 and 2. The percentage of the number of discarding chip required to avoid the out of spec after the screening can be increased by 5 orders of magnitude until the 15nm process generation compared to that of 40nm, as shown in Fig. 2. It has been indicated that almost chips have to be discarded around the 15nm process generation to avoid the out of spec in the market unless adequately treated with the MRASST designs.

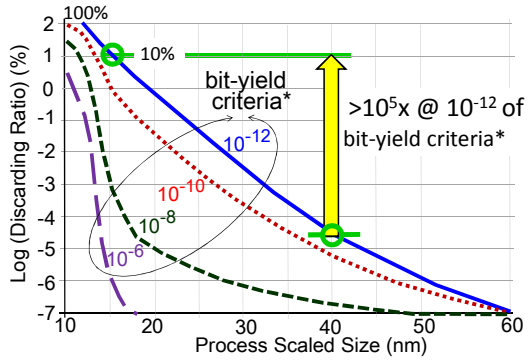

Fig. 2 Increased chip discarding ratio for 15nm can be a 10^5 x larger than that for 40nm.
Bit-yield criteria* is defined as the required fail probability, *i.e.*, 10^{-12} is for only 1-bit fail for 99.9% yield of 1Gbit chip

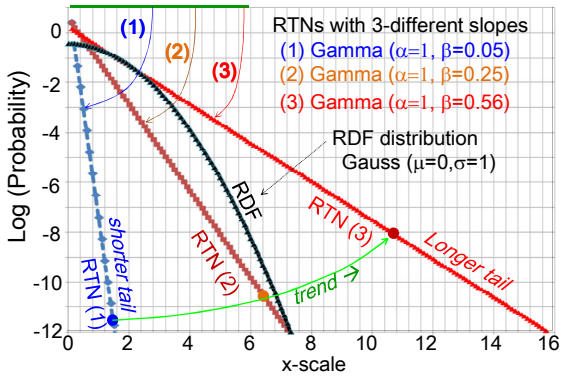

Fig. 3 Relationships between the tail distributions for RDF (Gauss) and three RTN (1), RTN (2), and RTN (3) (Gamma of β=0.05, 0.25, and 0.56, respectively)

In order to discuss the impacts of rapidly increased the TD-MV on the GB designs, the three cases of the ΔVth ratios of RTN/RDF: (1/4, 1/1, 4/1) are assumed in this discussion, as shown in Fig. 3.

Marked (1), (2), and (3) in Fig. 4 represent for the three cases of the relationship of the ratios of RTN/RDF, respectively. Here, we also assumed the three cases for RDF: RDF1, RDF2, and RDF3. These trends also make differently impacts on the trend of the RTN/RDF ratios. Since the advanced CMOS device tends to change to much less-dopant body devices like FinFET, ultra-thin body SOI, and nano-wire FET, there is the potential that the increasing paces of RDF are varied between 1/0.7, 1/0.84, 1/1 for RDF1, RDF2 and RDF3, respectively if assumed the LW is scaled down to 0.5 every process generation, as shown in Fig. 4.

In this paper, the yield-loss impacts made by the approximation-errors of the complex RTN distribution by various statistical models are discussed while considering the trend of the RTN/RDF ratios in the following sections.

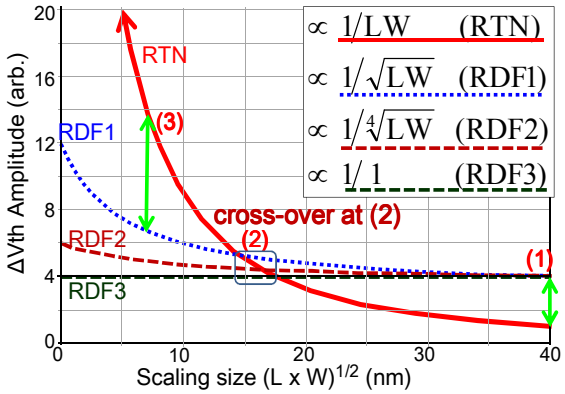

Fig. 4 Trends of ΔVth amplitude of RTN and RDF.
The increasing pace of RTN is assumed as proportional to 1/LW.
Three cases of increasing pace of RDF are assumed:1/0.7, 1/0.84, 1/1 for RDF1, 2 and 3, respectively

Here is how the rest of this paper is organized. In Section II, we discuss how the GB design for screening test will be changed when RTN becomes dominating over whole MV, followed by the impacts on the margin assist circuit designs to avoid the yield loss in Section III. In Section VI, we propose three types of fitting Gaussian mixtures model (GMM) following the discussions of challenges facing the conventional modeling in Section IV and V. In Section VII, we show the evidence indicating if the proposed models can approximate well the heavy long-tailed distributions and can give a precise fail-bit count prediction. We rigorously prove that it is possible to approximate more complex long-tailed distributions by mixtures of Gaussian distributions in Section VIII. Finally, we state our conclusion in Section IX.

II. Discussions on Issues of Guard Band Designs

The effects of long tail distributions on the shifted *screening point* (SP) are shown in Figs. 5-7. The different RTN

amplitudes of RTN(1), RTN(2), and RTN(3) are assumed, respectively.

The tails of *non*-TD-MV distributions by RDF are truncated by the screening. Additional tails are added after the screening by RTN caused TD-MV effects, as shown in Figs. 5-7. The convolution results of the two distributions of the truncated RDF and RTN show that there is the potential of the significant changes of the whole margin distributions in 10-years after the screening unless adequately treated with the MRASST designs.

As can be seen in the Figs. 5-7, the shallower-angled slope of the RTN distribution makes the length of tail longer. A longer tail makes the screening point more shifted (Δx). As shown in Figs. 5-7, the Δx for RTN(1), RTN(2), and RTN(3) are about 1, 7, and 10, respectively. In these examples, the screening point are assumed as x=-6, where the Gaussian distributions of RDF are truncated. It is worth mentioning that the impacts of the truncated distributions on the convolution results depend on the RTN slope. If the slope of RTN is steeper than the Gaussian RDF (case of RTN(1)), the distribution of the convolution results has a folding point like Q shown in Fig. 5. In contrast, the convolution for RTN(2) and RTN(3) does look like no effects on any truncated points. This is because the slope of RTN is shallower-angled than that of Gaussian. This indicates that any truncation of RDF can't control the tails of the convolution results any more.

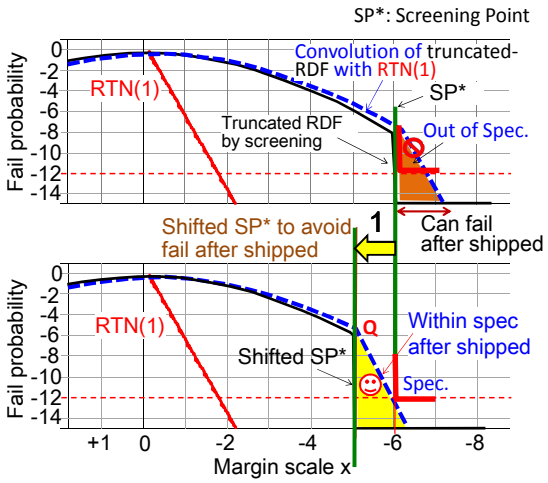

Fig. 5 RTN (1) impacts on the tails after the screening. To avoid any fail after shipped, the screening point has to be shifted by 1 of x. This causes additional chip discarding

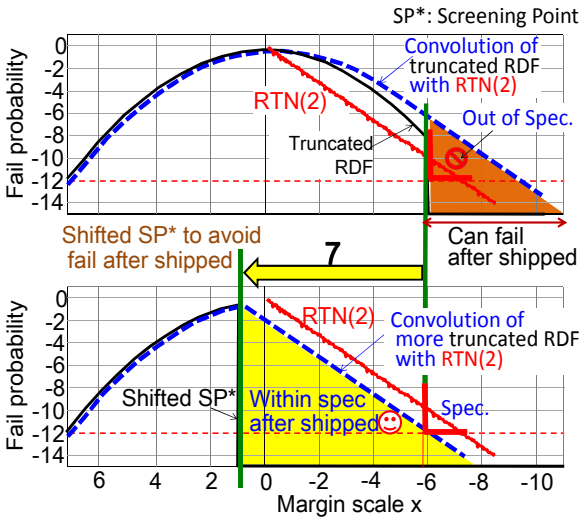

Fig. 6 RTN (2) impacts on the tails after the screening. To avoid any fail after shipped, screening point has to be shifted by 7 of x. This results in additional chip discarding

It is also worth noting that reducing the error of approximations to the RTN long-tail distribution is crucial challenge in the GB design. This is because the tail of the convolution probability density function (pdf) is strongly impacted by the tail of the RTN distribution.

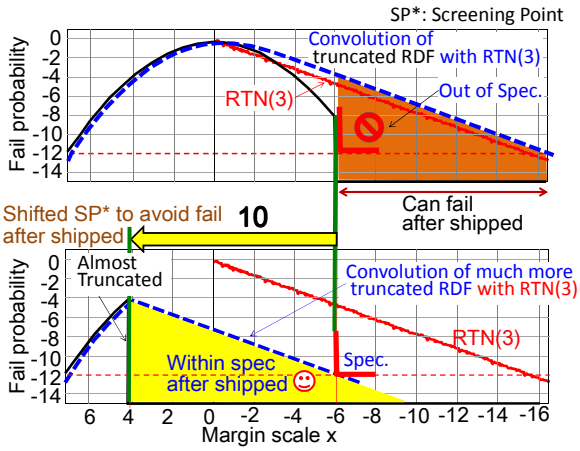

Fig. 7 RTN (3) impacts on the tails after the screening. To avoid any fail after shipped, screening point has to be shifted by 10 of x. This discards almost of the chips before shipped

The effects of the excessive chip-discarding yield-loss made by the error of the RTN approximation are shown in Fig. 8. More rarely event-analysis like its cdf $<10^{-12}$ requires a higher accuracy at a longer tail position (larger x) and its required error level depends on the interest cdf values. Thus, its errors in the three ranges of $10^{-(12\sim10)}$, $10^{-(10\sim8)}$, $10^{-(8\sim6)}$ are measured in this paper, as shown in Fig. 8.

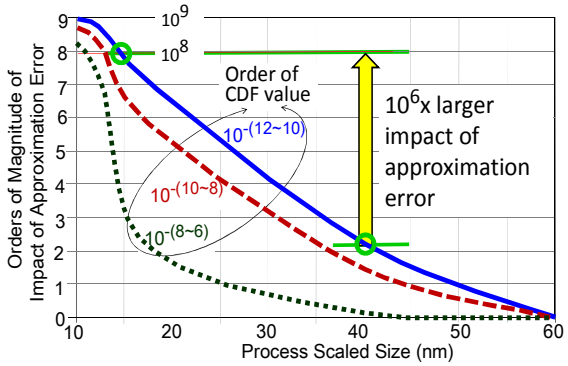

Fig. 8 Increased impact of approximation error on the trouble of the excessive under-estimation/over- estimation of the yield. It depends on both of the process scaled size and the order of cdf

It is worth mentioning that the accuracy of the RTN approximation by the statistical model will become more important as the process is scaled down. It stems from the facts that the RTN tail distributions will be longer and heavier due to the device size LW scaling, as shown in Fig. 4. As can be seen in Fig. 8, the errors affecting the discarding chip counts at 15nm can be over 6-orders of magnitude larger than that at 40nm in the cdf range of $10^{-(12\sim10)}$.

As explained in this section, the accuracy of the approximation of the RTN distributions is unprecedentedly crucial challenge for the GB designs to avoid an excessive under-estimation/over- estimation of the yield.

III. Assisted Margin Shifts

There are two potential means to avoid any out of spec after the screening, as shown in Figs. 5-7: (1) pre-truncating the less-margin chips so that "out of spec" never happens with the RTN-caused margin shifts ,as shown in Fig. 9 (left) and (2) increasing the margin by using the margin assist circuits (MRASST) [1]-[3] so that the convolution results of RTN and RDF can be fit within the spec, as shown in Fig. 9 (right).

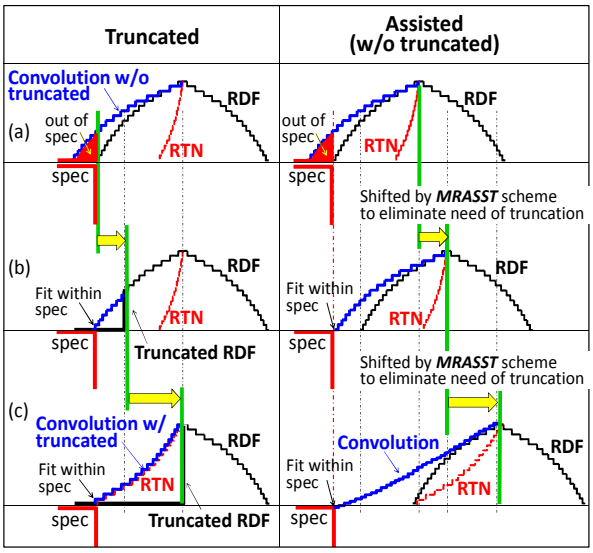

Fig. 9 Comparisons of the GB designs to avoid any out of specs after screening: (left) pre-truncating less margin chips so that out of spec can't happen with margin-shift by RTN and (right) increasing margin by using the MRASST [1] such that convolution results can be fitted within spec

As discussed in Figs. 5-7, the yield-loss will become prominent by the screening test unless adequately treated at the GB designs. To address this challenge, we will need to design the *margin assist circuits* (MRASST) for the RTN-caused variations as shown in Fig. 9 (right), which are conventionally used for the *non*-TD GB designs for the RDF-caused variations [1]-[3]. However, the conventional statistical models based on the Gaussian distributions can't be used for the MRASST designs any more to compensate the SRAM margin shifts by the TD-RTN caused variations. It stems from the changes of the statistical distribution of the whole MV from the simple Gaussian to complex Gamma mixtures distributions. In order to address the issues, the new models that provide large enough accuracy for the both distributions of Gaussians and *non*-Gaussian like Gamma mixtures are discussed in the following section.

IV. Challenge for Modeling of RTN Gamma Mixtures Distributions

According to [4]-[6], the distribution of the RTN amplitude will have the complex bounded tails caused by "atomistic" variation-behaviors with the various variation factors of the *gate line-edge roughness* (GER), the *fin-edge roughness* (FER), and the *metal gate granularity* (MGG) [4]-[6], 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. 10.

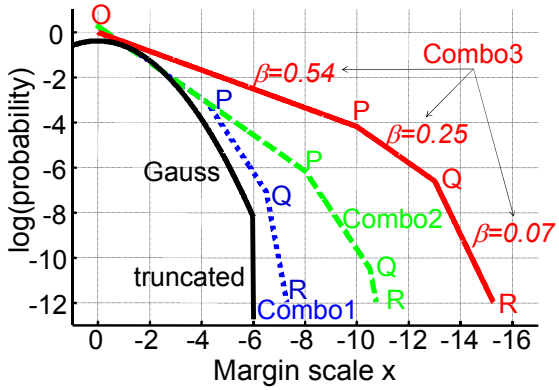

Fig. 10 Various bounded tails of the distributions of Gauss (RDF) and combination of different shaped gamma distributions of Combo1, Combo2, and Combo3

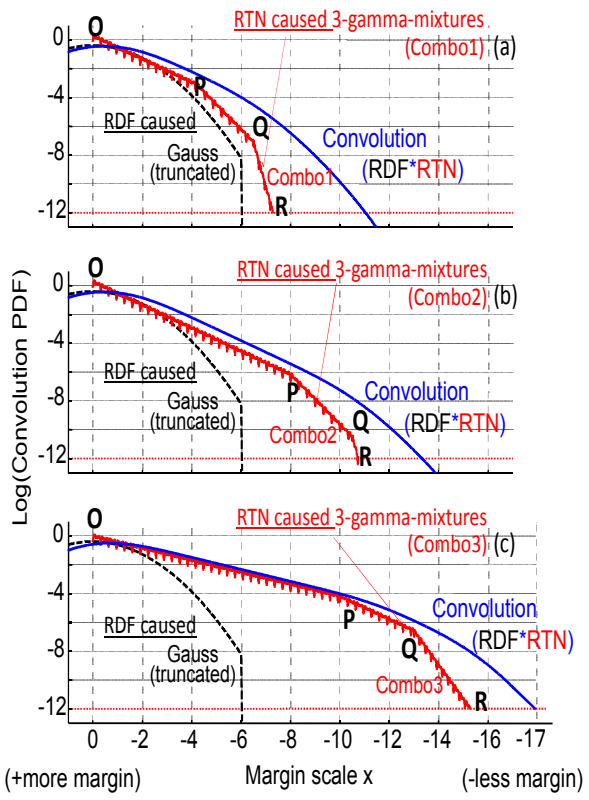

Fig. 11 Comparisons of the convolution results between the truncated Gaussian RDF and 3-different gamma RTN distributions: (a)combo1,(b)combo2, and (c) combo3

Fig. 11 illustrates the *probability density functions* (pdf) for the 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 the Gamma long-tail distributions of the RTN.

However, the appropriate approximation method for meeting the requirements for this application have not been proposed yet.

V. Issues of the Conventional Models

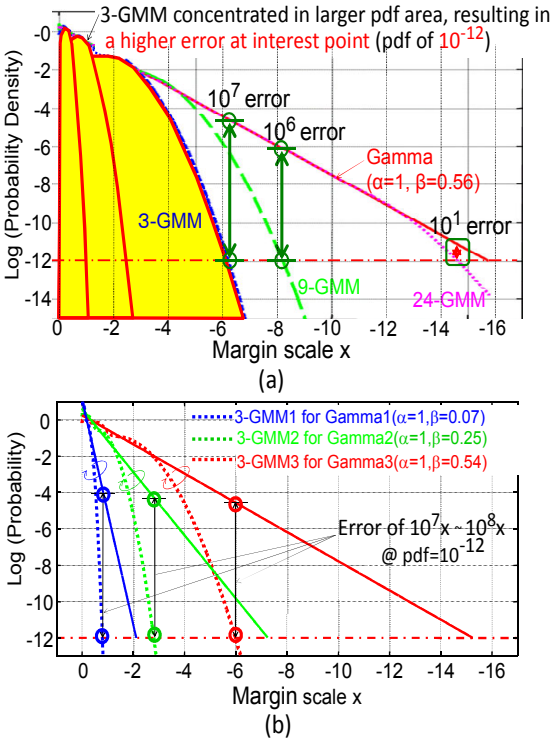

Fig. 12 (a) Approximation-error comparisons between 3, 9, 24-GMMs cases: errors of orders of 10^7 , 10^6, and 10^1, respectively. (b) β dependency of the coventional model errors for the 3 types of Gamma distributions of β=0.07, 0.25 and 0.54, respectively

The expectation-maximization (EM) algorithm [9], which is an iterative procedure that maximizes the likelihood of the Gaussian mixtures models (GMM), is well known as easy and convenient means to approximate the GMM to the *non* Gaussian distributions.

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. 12. 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. 12(b). Even if increasing the number of the 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. 12(a). In almost all FBC analyses, the distribution of interest only matters in the tail-region of the probability of orders of 10^{-12}

[6]-[8]. Thus, this is a crucial challenge facing the rare-event SRAM yield predictions. We should solve this issue until the time comes.

VI. Proposed Statistical Approximation Model for RTN Gamma Mixtures Distribution

In order to solve the above issues, we propose, for the first time, the three kind of fitting methods to approximate any arbitrary long-tailed RTN distribution by an adaptive segmentation *Gaussian mixtures model* (GMM).

These provide the following benefits: (1) applicable to the various convex and concave shapes of the bounded Gamma distribution even with the wide range of the 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 proposed fitting mixtures of the 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. These can be expressed by the analytical expressions, which allow using the normal (Gaussian) cumulative density function (normcdf) for estimating the error counts. This can give us the FBC by just summing up the values of the normcdf for each Gaussian of the whole GMM. The example of how to caluculate the the FBC of the segmentation of (x_a-x_b) is shown in Fig. 13. This makes it easier to predict the FBC before and after the screening at the stages of both circuit design and screening test.

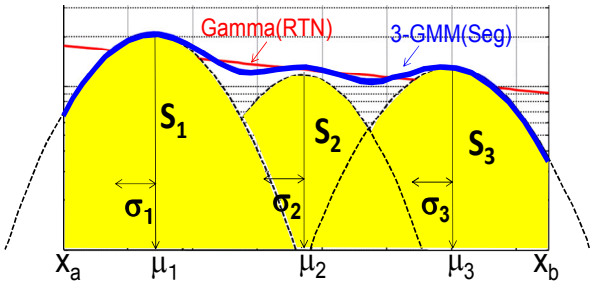

Fig. 13 Error bit counts of the segmentation of (x_a-x_b) can be given by just summing up the normal (Gaussian) cumulative density function (normcdf) of three GMMs

The centerpiece of these 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 the 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. 14(a)-(b) and Fig. 15, respectively.

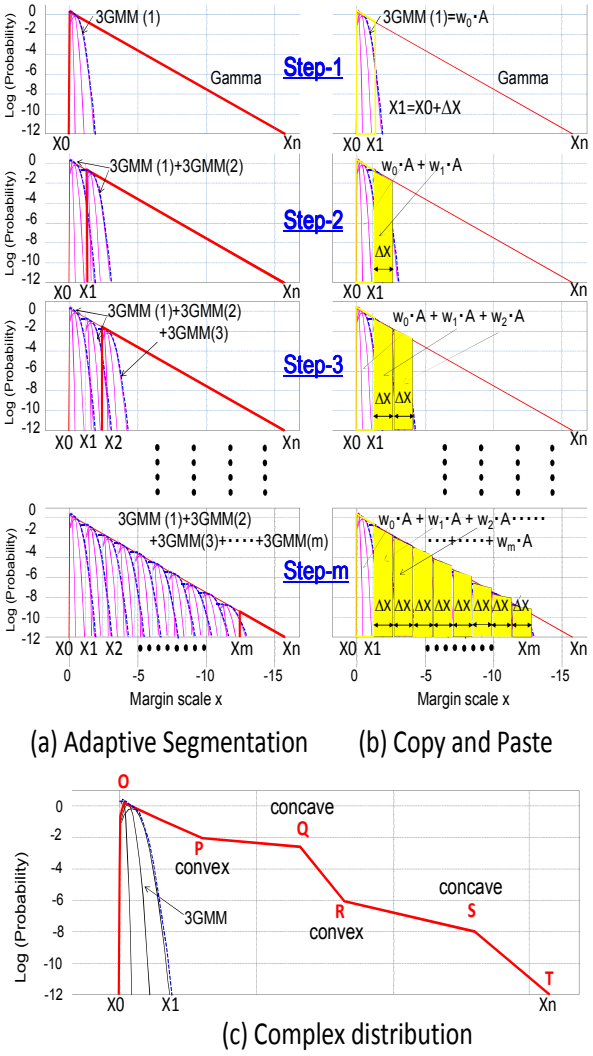

Fig. 14 Concepts of the proposed approximation algorithm. (a) adaptive segmenttaion:Xm are decided such that likelifood of each 3-GMM can be maximized. (b) copy and paste fashion: copy the first 3-GMM and paste to others with adaptive weighting. (c) example of complex distributions comprising various variation factors

A. Adaptive segmentation based fitting

Algorithm of the adaptive segmentation is described below from the step-(1) to step-(3).

(1) 1st-step is to do approximation by the 3-GMM between X0 and Xn. And find the point of X1, where the 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 the likelihood of the 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 the likelihood of the 3-GMM is maximized between X3 and Xn.

This flow can be repeated until the likelihood of the whole GMM can be maximized as shown in Fig. 14(a).

B. Copy and paste fashion based fitting

Algorithm of the copy and paste fashion is described below from step-(1) to step-(3).

(1) 1st-step is to do approximation by 3-GMM between X0 and Xn. And find the point of X1, where the likelihood of the 3-GMM is maximized. ΔX is given by (X1-X0) and w_0 is the weight of the 1st 3-GMM.

(2) 2nd-step is to get the weight (w_1) of the 2nd 3-GMM. And copy the 1st 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) 3rd-step is to do the same thing as (2), as shown in Fig. 14(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. 13(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

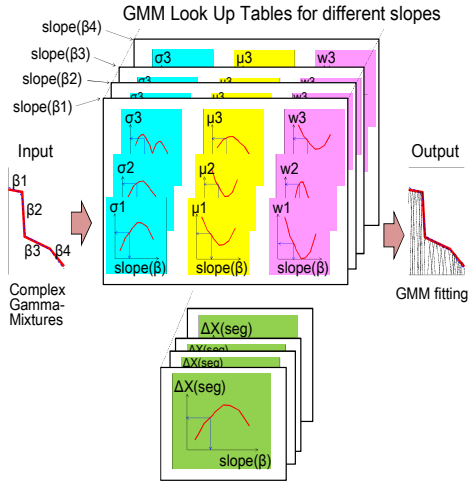

Fig. 15 Concept of the look up table (LUT) for the different sloped-gamma distributions of the shape parameter β. This LUT provides the parameter set of 3-GMMs (α,β,w) and width of segmentation which maximizes its likelihood in the segmentation

However, as the number of the folding points is increased, the number of EM operations required to get the GMM for the individual segmentation is also 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 the EM operations. If the information of the slope of the individual segmentation, e.g., β of shaped parameter of the gamma distribution is just input, the LUT outputs the all parameters required to regenerate the GMM comprising the 3-Gaussians, as shown in Fig. 15.

This also outputs the best width of individual segmentation ΔX(seg) that the likelihood can be made maximized.

As a result, overall approximations with the optimized segmentation width can be easily done without any time-consuming EM steps.

In this paper, we assumed the range of the slope is β=0.05～0.95, which corresponds to the variations of the slope of the log-scaled gamma distributions, as shown in Fig. 16.

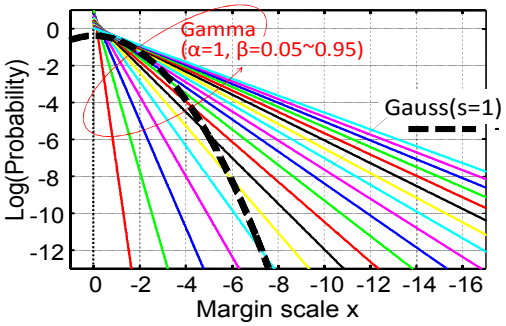

Fig. 16 Various sloped-gamma RTN distributions compared with the Gaussian distribution (σ=1) of RDF

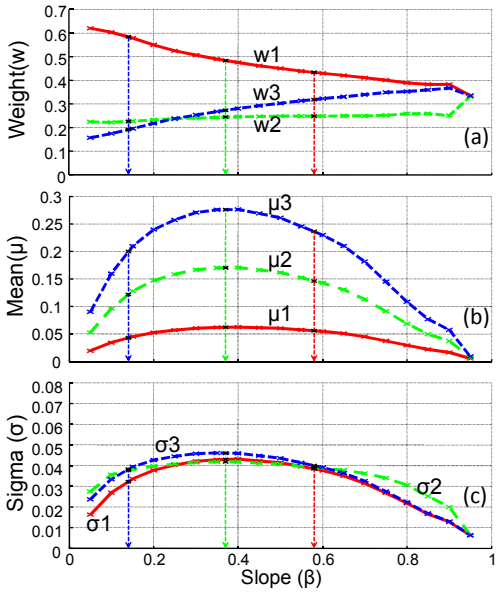

Fig. 17 Slope dependency (β of gamma) of the parameters being used in LUT for the three Gaussian mixture model (3-GMM)

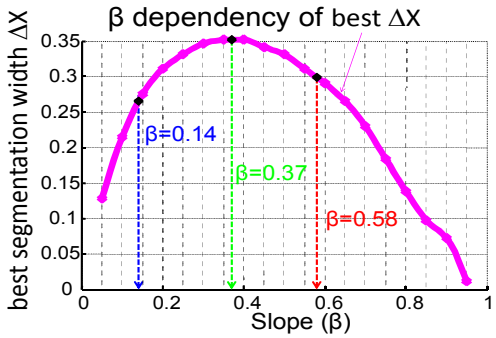

Fig. 18 Slope dependency of parameters in the LUT for the best segmentation width ΔX

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. 17 and 18, the error caused by interpolation of LUT can be minimized.

Fig. 19 shows the positions of the maximum likelihood and the minimum of the approximation error in the 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. 19. 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. 17).

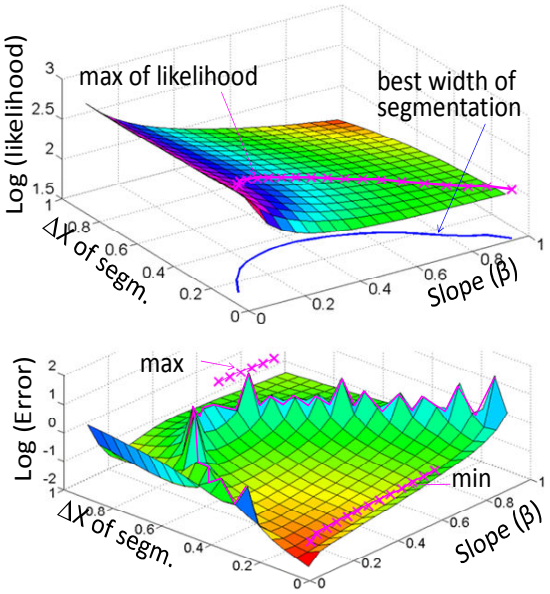

Fig. 19 Likelihood and approximation error dependencies of the slope-β and the segmentation width-ΔX

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

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. 20.

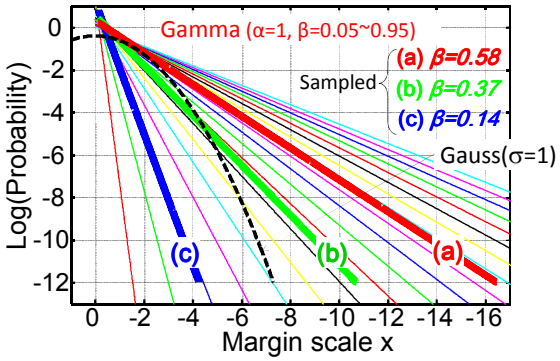

Fig. 20 Comparisons of the slopes of the tails between the Gauss for RDF and 3-sampling points of slope β=0.14, 0.37, and 0.58 for RTN

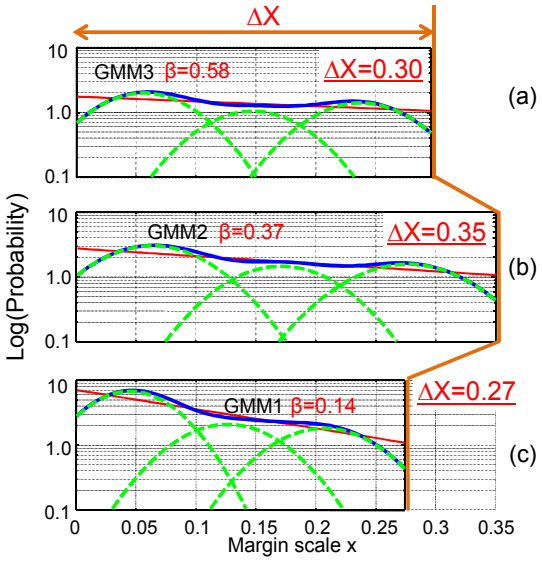

Fig. 21 3-GMMs in the different best ΔX-segmentation for different sloped tails of (a)ΔX=0.3,β=0.58, (b) ΔX=0.35, β=0.37, (c) ΔX=0.27,β=0.14

Fig. 21 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. 22 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].

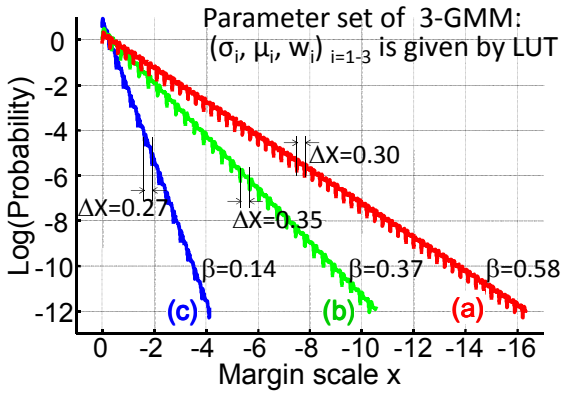

Fig. 22 LUT based fitting of the different sloped tails of (a)ΔX=0.3,β=0.58, (b) ΔX=0.35, β=0.37, (c) ΔX=0.27,β=0.14

Fig. 23 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. 23. 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.

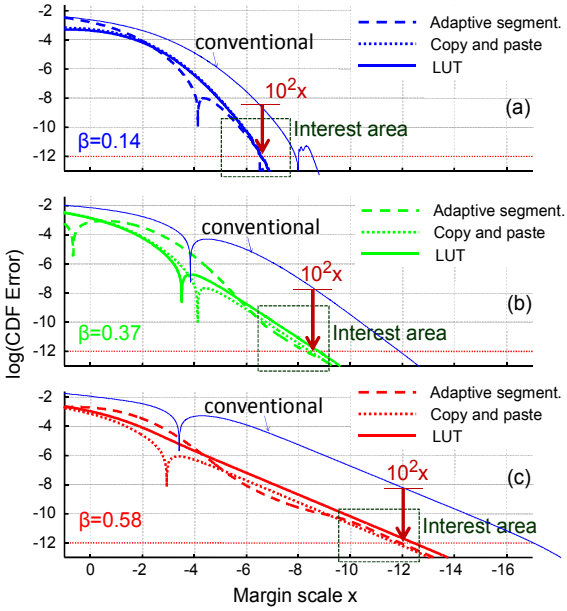

Fig. 23 Cdf error of the convolution results between Gauss (σ=1) and 3-different Gammas of (a) β=0.14, (b) β=0.37, and (c) β=0.58, respectively

VIII. Application to More Complex Distributions

According to [3]-[8], 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 Fig. 24. As its examples, the three types of distributions whose have a different slopes and folding points are assumed as Combo1, Combo2 and Combo3, as shown in Fig. 24.

The approximation-errors for fitting to Combo1, Combo2, and Combo3 are compared between the cases of using (a) the convnetional 3-GMM model and (b) the proposed segmentation models. As can be seen in the Fig. 24(a), the conventional 3-GMM models without using segmentation manner can't fit the tails of Combo1-3 at all. The errors of 4,6, and 7 orders of maginitude have to be expected at the rare probability of 10^{-12}. Contrary, the fitting errors can be drastically reduced by using the proposed ideas, as shown in Fig. 24(b). Unlike the case of Fig. 24(a), it can be seen that the fitting curves and its target lines in Fig. 24(b) are perfectly overlapped. Thanks to the segmentation manner, the same concepts can be adaptively applied to the different sloped-tail distributions. This indicates that this ideas can be applied to the various sloped-distributions even if they are combined like the given examples in Fig. 10.

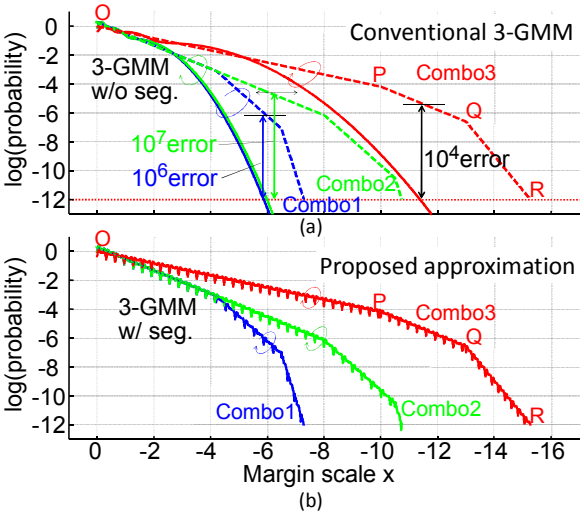

Fig. 24 Comparisons of approximation-errors for fitting to Combo1, Combo2, and Combo3 between the cases of (a) with the convnetional 3-GMM model and (b) with the proposed segmentation models

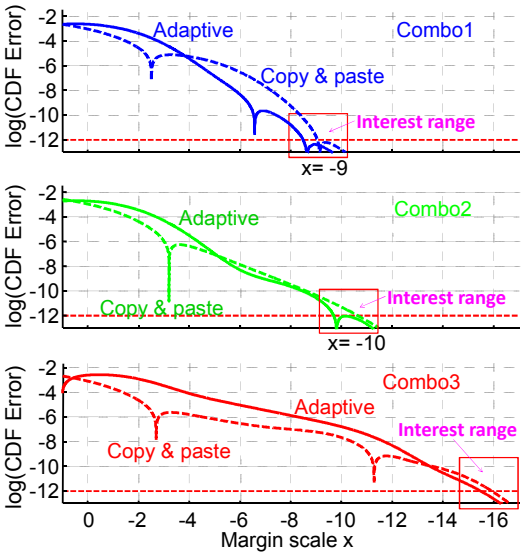

Fig. 25 Comparisons of the errors of cumulative density function (cdf) of the convolution results for Combo1, Combo2, and Combo3 between the case of using the "adaptive segmentation" and "copy and paste" fashion

Since the both ideas of "adaptive segmentation" and "copy and paste" fashion can apply to this kind of complex non-linear distribution, the errors of cumulative density function (cdf) of the convolution results for Combo1, Combo2, and Combo3 are compared between the two, as shown in Fig. 25.

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. 25.

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.

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. 26(a). 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 [3]-[5] uses such kind of shapes as an example of the potential future RTN distribution, as shown in Fig. 26(b).

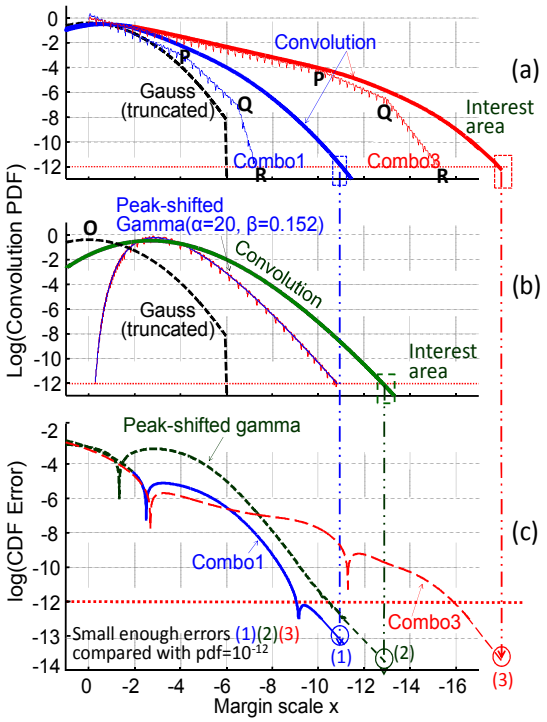

Fig. 26 Convolutions of (a) 3-different sloped combined gamma distributions and the truncated Gauss distribution and (b) peak-shifted gamma. (c) Cdf-error comparisons between the three different distributions of (a) and (b)

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. 26(a) and (b) shows, while reducing the error of cdf to less than 1%, as shown in Fig. 26(c). As can be seen in Fig. 26, the 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.

IX. Conclusion

This paper proposes, for the first time, how the challenges facing the GB designs including the MRASST schemes for the screening-test in the coming process generations should be addressed. It has been shown that yield-loss (chip-discarding) by screening test may become crucial issues if RTN could not be reduced or eliminated. It has been pointed out that intolerable yield-loss by wrong GB design can be increased by 6-orders of magnitude. The required accuracy of statistical model for approximating the tails of RTN distributions will

become unprecedentedly crucial as the process is approaching to a 15nm and beyond.

In this paper, we have proposed, for the first time, the three types of GMM fitting schemes for approximating the complex gamma mixtures which are combination of the various-sloped distributions with multiple convex and concave folding points. We show that how much its approximation-error can affect on the accuracy of the statistical predictions of the FBC, which is required to avoid the out of spec after shipped to the market. It has been pointed out that proposed fitting methods can provide the practical fitting models to predict the failure probability during the life-time due to the time-dependent RTN-caused failures. This can adapt any arbitrary sloped mixtures distributions without any need of computing power for the EM convergence.

It has been verified that the proposed three types of methods can reduce the error of the FBC predictions by about 4-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.

We have pointed out that the proposed methods are one of candidate fitting algorithms, which will be crucial not only for the SRAM GB design but also the MRASST design in the coming process generations.

ACKNOWLEDGMENT

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

REFERENCES

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

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

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

[4] X. Wang, A.R.Brown, B.Cheng and A.Asenov "RTS amplitude distribution in 20nm SOI FinFETs subject to Statistical Variability", *SISPAD* 2012, pp.296-299

[5] X. Wang, G.Roy, O.Saxod, A.Bajolet and A.Juge, A.Asenov "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

[6] K.P.Cheung, J.P.Campbell, S.Potbhare and A.Oates " The amplitude of random telegraph noise: Scaling implications", *Reliability Physics Symposium (IRPS), 2012 IEEE International*, pp.1.1 - 1.3

[7] K. Takeuchi, T.Nagumo and T.Hase "Comprehensive SRAM Design Methodology for RTN Reliability" , *Digest of IEEE Symposium on VLSI Technology*, (2011), pp. 130-131

[8] K. Takeuchi, T.Nagumo, K.Takeda, and S.Asayama, S.Yokogawa, K.Imai, K.Hayashi "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

[9] Moon, T.K, "The expectation-maximization algorithm" in *Signal Processing Magazine, IEEE*, Volume: 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 (M'98) 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 machine learning based 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.