FIT Fukuoka Institute of Technology 福岡工業大学

GRADUATE SCHOOL OF ENGINEERING

FUKUOKA INSTITUTE OF TECHNOLOGY

A Study on Time Dependent Variations Caused Rare-Event Fail Bit Count Prediction Analysis Methods for Nano-Meter Scaled SRAM Reliability Designs

Worawit Somha

Adviser: Prof. Hiroyuki Yamauchi

2014

Contents

List of Figures

List of Tables

List of Abbreviation and Symbols

The abbreviations appearing in the dissertation are listed below. Any minor departure from these abbreviations is explained in the text itself.

Abbreviation	Description
ASSTS	assisted circuit schemes
cdf	cumulative distribution function
CMOS	complementary metal oxide semiconductor
DIAS-BDCV	iteration loop and an adaptively segmented forward-problem based blind deconvolution
EM	expectation maximization
FBC	fail bit counts
FER	fin edge roughness
GMM	Gaussian mixtures models
GER	gate line edge roughness
GB	guard band
IOPD	iterative optimization problem based deconvolution
IOPLSD	iterative optimization problem based deconvolution with least square
ISDCN	iteration loop segmented forward problem based deconvolution
IASDCN	iteration loop and adaptively segmented forward-problem based deconvolution
LUT	lookup table
LAD	linear algebraic deconvolution
L/W	length and width of transistor
MRASST	margin assist circuits
MGG	metal gate granularity
MOSFETs	metal oxide semiconductor field effect transistor
MV	margin variations
NBTI	negative bias temperature instability
PBTI	positive bias temperature instability

Abbreviation	Description
PFDCV	proposed partitioned forward problem based deconvolution
pdf	probability density functions
RDF	random dopant fluctuation
RTN	random telegraph noise
R-L	Richardson-Lucy
SoC	system on chip
SRAM	static random access memory
SP	screening point
TP	truncating point
TD	time dependent
V_{th}	threshold voltage
$V_{DD,min}$	minimum operating voltage

Greek symbols

μ	Mean of Gaussian distribution
σ	Sigma of Gaussian distribution
α	Locator parameter of gamma distribution
β	Shape parameter of gamma distribution

Abstract

At the first, this study proposes a fitting method to approximate the mixtures of various sloped-tail Gamma distributions 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 Expectation-Maximization (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.

Then, this study also proposes 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 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 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.

Following that, a lookup table (LUT) based fitting is newly proposed method to approximate the mixtures of various sloped-Gamma tail distributions by an adaptive segmentation 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.

Then, this study also discusses, for the first time, how the statistical SRAM design analyses should be changed when: 1) the shift-amount of the time-dependent (TD) voltage margin variations (MV) after the screening test will become larger than that before and 2) the shapes of the MV distribution will change from the Gaussian to the complex mixtures of Gamma distributions. We discuss on the SRAM TD-MV analyses with not only the forward problem but also the inverse problem, i.e., deconvolution analyses. The proposed algorithm for the deconvolution to circumvent the issues caused by high-pass filtering behavior is discussed. Based on the proposed convolution/deconvolution design analyses, it has been shown for the first time that: 1) detecting the truncating point of the distributions of TD-MV by the screening test and 2) predicting the required the MV-shift-amount by the assisted circuit schemes to avoid the out of specs in the market during the life-time, etc, has become enabled based on the target specification.

This study makes a comparative review of methods for the SRAM deconvolution analyses and discusses the pros and cons between the two methods of linear algebraic deconvolution (LAD) and the iterative optimization problem based deconvolution (IOPD) to extract the random dopant fluctuation (RDF) and/or random telegraph noise (RTN) from the complex mixtures of gauss and gamma distributions characterizing an overall nano-scaled SRAM margin variation. We have demonstrated and shown for the first time that 1) the IOPD method can avoid the abnormal errors in both cases for the RDF and RTN deconvolution and 2) the LAD method can suppress the deconvolution error for the RDF more than IOPD but unfortunately confronts the issues on excessive deconvolution errors for the RTN.

This study proposes a ringing-error-free non-blind deconvolution technique featuring an iterative and adaptively segmented forward-problem based deconvolution (IASDCN) process. Unlike the algebraic based inverse operations, this eliminates any operations of differential and division by zero to successfully circumvent the issue on an abnormal V-shaped error. This effectiveness has been demonstrated for the first time with applying to a real analysis for the effects of the RTN and/or RDF on the overall SRAM margin variations. It has been shown that the proposed IASDCN technique can reduce its relative errors of RTN deconvolution by 1013 to 1015 times, which are good enough for avoiding the abnormal ringing errors in the RTN deconvolution process. This enables to suppress the cdf error of the convolution of RTN with RDF (i.e., fail-bit-count error) to $1/10^{10}$ error for the conventional algorithm.

This study extends the proposed ideas to a different application of blind deconvolution technique for extracting the unknown two variation factors of the RTN and the truncated RDF solely from the given target for SRAM margin variations. Unlike the non-blind

deconvolution, the blind deconvolution has to extract the both of the two unknown factors of RTN and truncated RDF simultaneously, that can be sort of ill-posed problem. The proposed algorithm features a sequentially-dual iteration loop and an adaptively segmented forward-problem based blind deconvolution (DIAS-BDCV) process. This allows a free of convergence error in the optimization process. This effectiveness has been demonstrated for the first time with applying to a real SRAM design analysis. It has been shown that the proposed DIAS-BDCV technique allows: 1) a free of convergence-error and local-minimum-error in blind deconvolution even if the total number of parameters to be sought in the optimization problem exceeds 20, and 2) a low enough blind deconvolution errors of the RTN and RDF comparable to the level ($< 10^{-13}$) of the non-blind one.

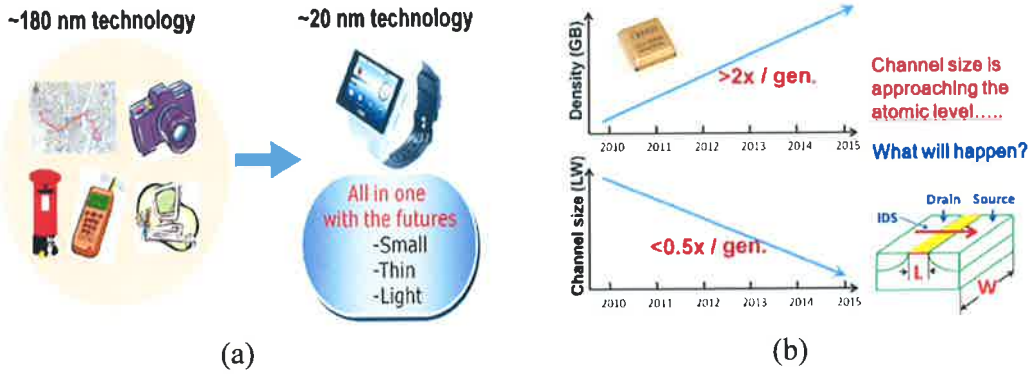
In order to clarify the effectiveness of this study, a comparative analysis of the RTN deconvolution accuracy between the Richardson-Lucy (R-L) algorithm and the proposed partitioned forward problem based deconvolution means (PFDCV) are discussed. Unlike the R-L based deconvolution, the proposed technique successfully solves the issue of noise amplification thanks to eliminating any operations of differential and division. This effectiveness has been demonstrated for the first time with applying it to a real analysis for the effects of the RTN on the overall SRAM margin variations. It has been shown that the proposed PFDCV technique can reduce its relative errors of the RTN deconvolution by 10^{14}–fold compared with the cases of the R-L.

Chapter 1

Introduction

Following the explanations about the evolution of the computer system with the VLSI device size scaling, the critical issues of spatial and temporal frequency domain variations posed by device scaling are introduced first followed by discussing the objectives of this study in the Chapter-1. As a background for the objective of this study, the demanding trend for increasing memory density and its effects on the required accuracy level for the statistical analyses are also discussed first.

1.1 Background



(a) (b)

Figure 1.1: (a) All functionalities are integrated in one chip, (b) Trends of memory density and device size scaling based on Moore's law.

With the advent of Internet of Things (IoT) era, system-on-chip (SoC) is expected to become more and more critical part in the electronic world. The SoC is intensively being asked to be less power consumption, smaller and thinner, more multiple built-in functions and so on, in order to allow the evolution of the computer system, as shown in Fig. 1.1 (a). Continuing Moore's law [1] drives a continuing scaling down of the SoC and its vision has been made a reality during the last four decades. Thanks to its sustained CMOS scaling in four decades, the element device has been made smaller by orders of magnitude and enabled to pack over 10 billions of them on a single SoC. This allows us to design and ship

a high-performance commodity Smartphone to our society while meeting the above mentioned and exploiting its great benefits, as shown in Fig. 1.1(a).

The bit-density of the Static Random Access Memory (SRAM) in a single SoC is constantly increased by 2-times every 2-years thanks to sustainable size scaling down by 0.5-times every 2-years based on the Moore's law [1], as shown as in Figure 1.1 (b). As a result, the 1-bit fail-probability becomes smaller and smaller. In the following subsection, the relationship between the trend of the SRAM bit density and the requirements for the accuracy of the fail-probability prediction accuracy is discussed.

1.1.1 Ever-Increasing Memory Bit Demands

One of the main contributions of this study is to propose the approximation methods for unprecedentedly long tailed distributions, i.e., unprecedented rare event, with mixtures of Gaussian probability distributions. Since the Gamma distribution has a longer tail than Gaussian and the tail length can be expressed by a parameter β, the Gamma distribution is used to express the tail length in this study, as shown in Figure. 1.2.

In order to clarify the relationship among the three parameters: 1) memory bit density, 2) slope parameter β of Gamma distribution and 3) the tail length expressed by raw score x, the max values of raw score x to be considered, i.e., tail length, are plotted for four different memory density cases, as a function of the slope parameter β, as shown in Figure. 1.2. The raw score x corresponds to the number of sigma (σ), i.e., the number of Z for the Gaussian distribution.

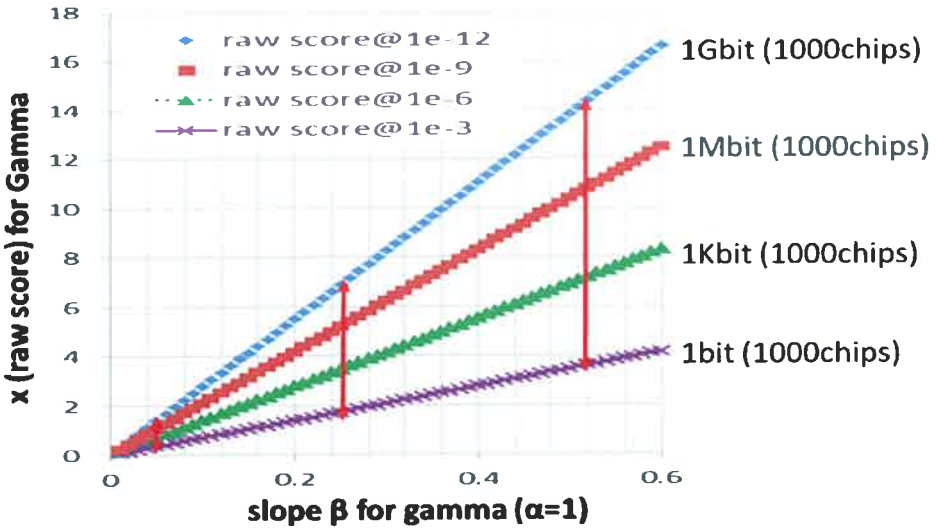


Figure 1.2: The relationship of three parameters of 1) memory bit density, 2) slope parameter β and 3) the tail length expressed by raw score x.

The four lines represent for the case of 1000 chips of 1-bit, 1K-bit, 1M-bit, and 1G-bit, respectively. This relationship indicates that: 1) ever increasing memory bit density extends the tail length x and 2) its tail length x depends on the slope β of the Gamma distribution. If the slope β is smaller, the x score range has less dependency on the memory density. On the other hand, as the slope β is increased, the memory bit density dependency of the raw score x becomes stronger, as can be seen in Figure. 1.2. According to the trend of the value of β, it is expected to become larger because the complexity of comprising elements of the whole distribution and the memory bit-density are increased. Thus, the raw score x to be considered will become further increased, as represented by the top line in Figure. 1.2.

One more important thing to be considered is the memory density dependency of 1-bit fail probability. For example, the 1-bit fail-probability for the 1000 chips of 1-bit, 1K-bit, 1M-bit, and 1G-bit memory are 10^{-3}, 10^{-6}, 10^{-9}, 10^{-12}, respectively. Thus our attention zone of this is 10^{-12}, which is unprecedentedly small number for the fail-probability to be considered.

The summary of this subsection is as follows:

1. Analysis of fail bit count (FBC) for larger number of memory bits requires more accuracy in a larger number of x (longer tail region).
2. Analysis of longer tail (larger β) distribution requires more accuracy in a larger number of x (longer tail region).
3. Growing trend toward to "Larger number of bits and longer tail distribution"

1.1.2 Increasing Complex Variation Factors

At present, MOSFET (Metal Oxide Semiconductor Field Effect Transistor) device is widely used as the key component of LSIs. As mentioned above, in order to continuously meet the requirements from the electronics industry, the scaling treadmill will have to sustain to fulfill the demand of the performance per cost for ever. However the size of element device has approached to an atomic size and conventionally ignored time-dependent (TD) atomic behavior cased variations will pose as a major challenge and barrier to a secure device screening design for avoiding any reliability troubles in the market.

This section describes two basic factors causing variation of the device characteristic distributions such as threshold voltage (V_{th}) : 1) Random Dopant Fluctuation (RDF) [2] that does have spatial dependency but doesn't have time dependency at all and 2) Random Telegraph Noise (RTN) [3,4,5] that does have both of the spatial and temporary dependencies.

1) Random Dopant Fluctuation (RDF)

During the fabrication of MOSFET, some impurity atoms (dopant) need to be doped to the silicon to adjust the value of V_{th}. However, the device size approaches to the dopant molecule size, the number of dopant and position of atoms in the limited channel area cannot be controlled any more as accurately as the designer want to achieve.

This indicates that no one can control those and random values for those parameters have to be accepted. This kind of randomness causes to randomly modulate the channel potential distribution in spatial frequency domain, as shown in Fig. 1.3. That is why the V_{th} has a random variation and its whole distribution obeys the Gaussian. As the size of channel is reduced, the randomness is increased, resulting in wider distribution with larger σ value, as shown in Fig. 1.3. In this study, it is assumed that the distribution of the V_{th} variation caused by the RDF effects obeys Gaussian distribution and expressed by σV_{th} and mean value $\mu = \mu_0$.

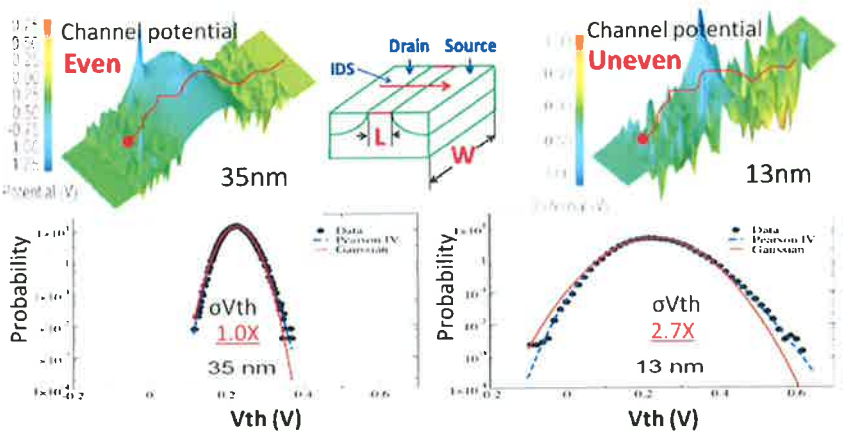


Figure 1.3: Comparison of the variation of the channel potential and V_{th} between the two device sizes for 35nm and 13nm.

2) Random Telegraph Noise (RTN)

After shipment of chip to the market, due to some external or internal factors that can change the energy of an electron, such as temperature, V_{DD} and so on, the electrons which stay normally at the channel can transit to another state in the energy band, such as a defect inside the oxide layer or interface area between the channel and the oxide. This phenomenon is referred to as RTN. While the single electron jumps into the oxide layer (trapped) from the channel, this state is defined as capture state, as shown as Figure 1.4. But while the single electron comes back to the channel (de-trapped), this state is defined as empty state. Along with the time dependent, the two states occur alternately, which leads to

the large variation of the value of V_{th}. The RDF and RTN variations are independent relationship thus the distributions of combined variations of RDF with RTN can be estimated by the convolution ($RDF_{VT\text{-}PDF} \otimes RTN_{VT\text{-}PDF}$) of the Gaussian distribution (σV_{th_RDF}, μV_{th_RDF}) for RDF with Gamma distribution (α, β) RTN, as shown in Fig. 1.5.

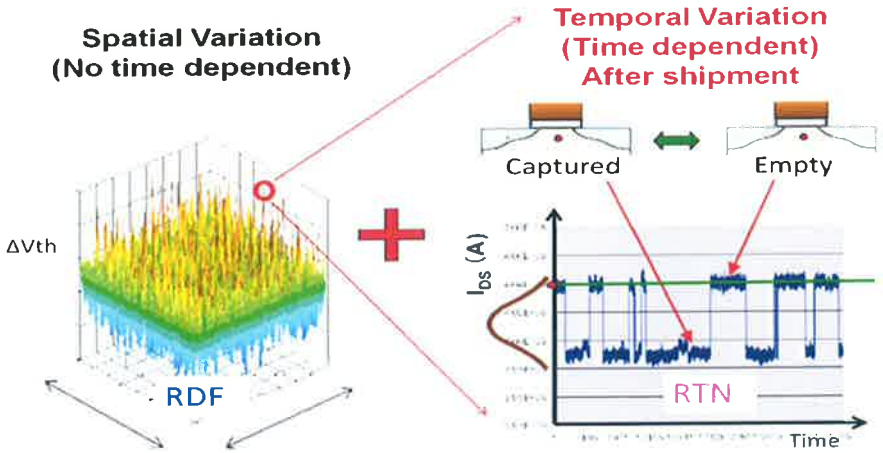


Figure 1.4: Effect on V_{th} variation with an increase in time.

The number of trapped electrons is integer but unknown, sometimes may be one, and sometimes may be two or three. If the quantity of the electrons in the channel is larger enough than trapped/de-trapped electron, its impact of trapped electrons can be ignored. However, as the size scaling is advanced, the impact on the V_{th} could not be simply ignored because the quantity of the electrons in the channel becomes less and less. Thus, in this study, both variations caused by RDF and RTN are considered simultaneously to predict the overall margins at 10-years later from the shipment, as shown in Figure 1.5.

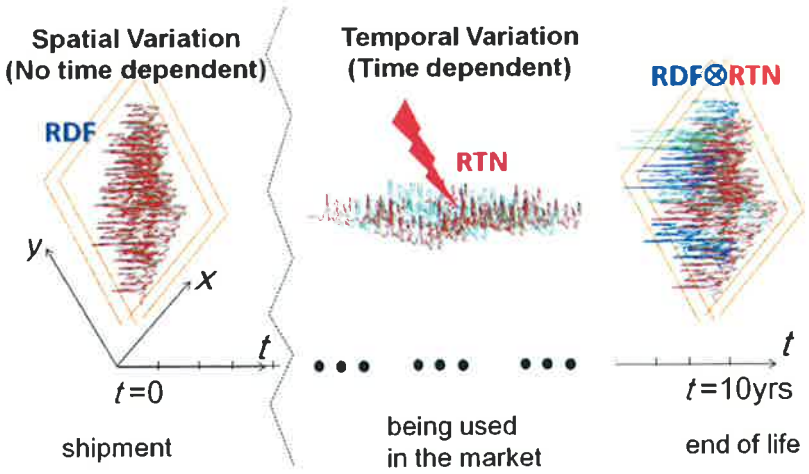


Figure 1.5: The RDF and RTN caused variations in both of the spatial and time frequency domains.

3) Comparison between RDF and RTN

Table 1.1 shows the comparisons between the RDF and RTN in terms of 1) when variation happens, 2) dependency of spatial and temporal frequencies, and 3) increasing pace with the channel are size that is defined by L x W. Herein, L and W are referred to as the channel length and width of the MOSFET [6,7], respectively. Some different characteristics can be seen between RDF and RTN effects, as shown in Table 1.1. The increased speed for the RDF and RTN effects on V_{th} variation has a direct proportional to the $1/(LW)^{0.5}$ and $1/(LW)^{1.0}$, respectively, as shown in Figure 1.6. This indicates that increasing pace of the amplitude of V_{th} variation for the RTN is 1.4-times faster than that for the RDF.

As shown in Figure 1.6, the RTN effect is conventionally much smaller than that for RDF at the timing point (1). However, its effect currently has been almost equal to that for RDF like the point of (2). For the future generation at point (3), the RTN effect is expected to become much larger than that for RDF owing to the 1.4-times faster increasing speed for the RTN effect. That is the reason why this study intensively considers the RTN effects other than that of RDF.

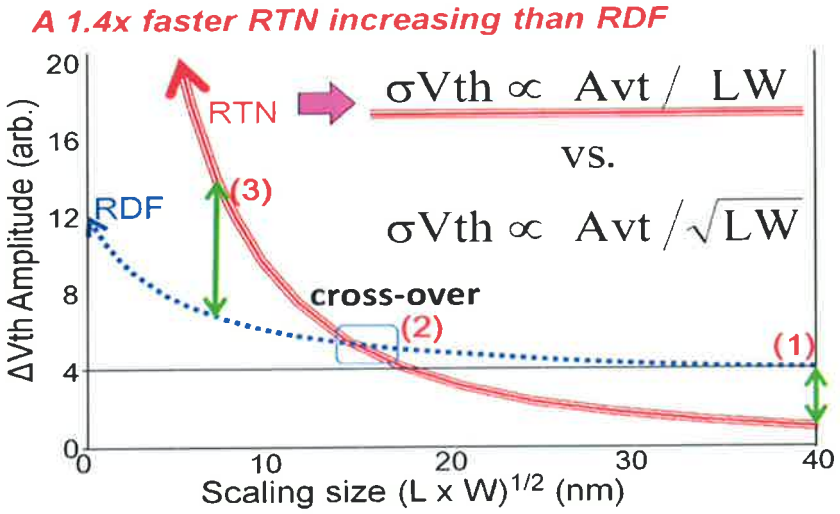


Figure 1.6: V_{th} variation increasing pace comparison between RDF and RTN.

Table 1.1 Different characteristics caused by RDF and RTN.

Name	Happening time	Dependency	Direct proportional
RDF	In Fabrication	Spatial frequency	$1/(LW)^{0.5}$
RTN	After shipment	Spatial and temporal frequency	$1/(LW)^{1.0}$

In this study, three sloped cases for RTN1, RTN2, and RTN3 are assumed to study the slope dependencies of the approximation accuracy and deconvolution errors, as shown in Figure 1.7. When comparing the tail length of the RTN1, RTN2, and RTN3 with the RDF, it is found that (1) RTN1<RDF, (2) RTN2=RDF, (3) RTN3>RDF, respectively. The corresponding process generations for RTN1, RTN2, and RTN3 are a 40nm, a 16nm, and a 7nm, respectively.

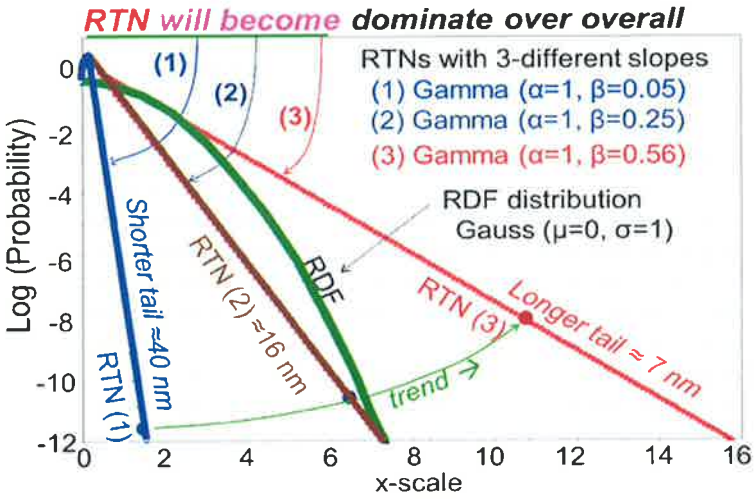


Figure 1.7: Comparisons between 3 cases of assuming RTN for 40nm, 16nm, and 7nm.

1.1.3 Increased Concern for Extra Time Dimension

Figure 1.8 illustrates the probability density functions (pdf) for RDF, RTN1(40nm) and RTN3 (<7nm), respectively. It is worth mentioning that the distribution-shape of the convolution result (f⊗g) obeys the Gaussian when RTN<RDF, and finally will become dominated by Gamma distribution of RTN when RTN>RDF, respectively.

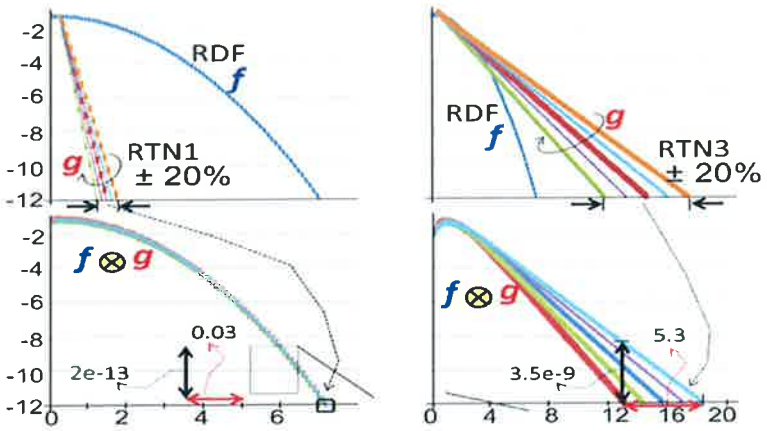


Figure 1.8: Comparisons of the impact of the approximation error of the RTN **g** on the tails.

Figure 1.8 shows the comparisons of the impact of the approximation error of the RTN **g** on the tails of the convolution results **h=f⊗g** between the 2-cases for RTN1 and RTN3. Based on the fact that the RTN tail length becomes larger than that for the RDF, the approximation-error of the tails of RTN distribution can be expected to become an unprecedentedly crucial challenge because of the following facts:

1. Probability density function (pdf) of 10^{-12} is touched at a larger x (raw score).
2. Gamma distribution requires a larger x compared with Gaussian because a derivative of Gamma distribution keeps the same value unlike Gaussian distribution.

It is found that the effects of a ±20% RTN **g** approximation error on the convolution (**h=f⊗g**) error can be ignored for the case of RTN1 but can be prominent for the case of RTN3, as shown in Fig. 1.8. The error level for the RTN3 is a quite different orders of the magnitude compared with the case of the RTN1.This is because the tails of **h=f⊗g** is no longer Gaussian but more governed by the RTN **g** for RTN3. The tail of **h=f⊗g** for the RTN1 is still Gaussian. This means that the reducing the error of approximations of the RTN **g** will become more crucial challenge in the screening test design as the length of the tail of the RTN **g** is increased.

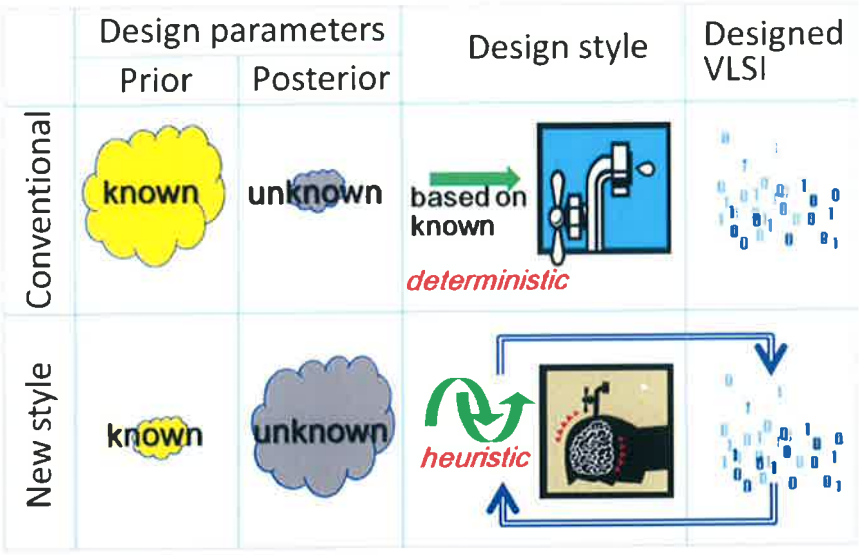


Figure 1.9: Proposed new VLSI design style.

The conventional VLSI design is based on known parameter and its design style is deterministic. On the other hand, from now on, the design style should be changed into new style because unknown parameter becomes larger than pre-defined parameter then the design style becomes heuristic (back and forth) as shown in Figure 1.9.

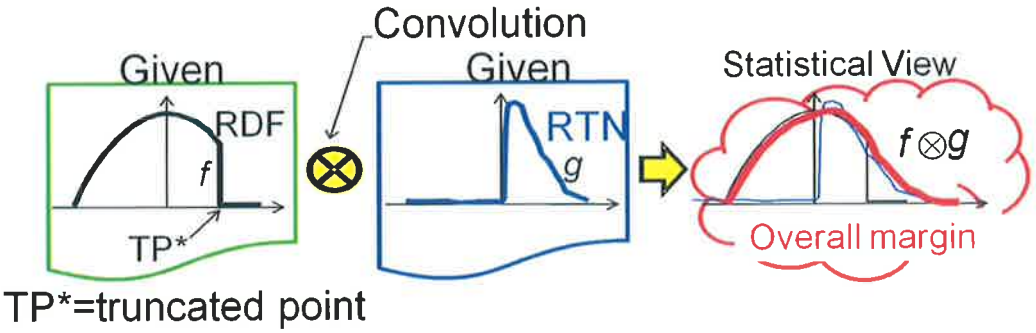


Figure 1.10: Fail-Bit-Count (FBC) prediction based on convolution RDF and RTN.

The traditional SRAM statistical analyses rely on the Gaussian model given its parameters extracted by the measured data. However, if the non-Gaussian unknown factors account for no longer just a fraction but a large percentage of the whole margin variation, this study has solved the non-Gaussian inverse problem based on the pre-defined hypothesis of the unknown factors or final target specifications. Figure 1.10 shows the concept for the required convolution methods to extrapolate the distribution for the whole margin variations based on the given data of the RDF and the RTN. Thus, precision of fail bit count (FBC) prediction relies on accuracy of approximation of tail of f (RDF), g (RTN) and accuracy of tail of h (=f⊗g).

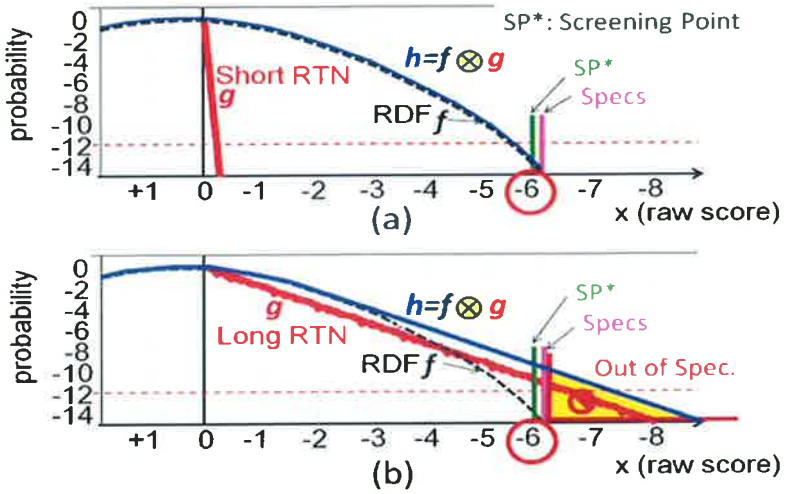


Figure 1.11: Comparisons of the RTN effects on the OVMV of (a) short RTN and (b) long RTN.

Figure 1.11 compares the dependencies of the RTN **g** on the convolution result **h=f⊗g** between the two cases of (a) short RTN and (b) long RTN. The overall voltage-margin variations (OVMV) is given by the convolution result **h** of the RDF **f** with the RTN **g**, i.e.,

h=f⊗g, as shown in Figure 1.11. As the tail length of the RTN **g** is increased, the distribution shape of the **h** becomes more governed by the RTN **g**, instead of the RDF **f**. This means that the shape of the distribution **h** becomes changed from the Gaussian to the Gamma distribution. As a result, ordinary Gaussian based margin analyses become no longer available. Since the cumulative density function (cdf) of the **h** corresponds to the fail-bit count (FBC), the approximation error of the **g** will have a direct impact on the precision of the FBC prediction.

This is the reason why the ordinary screening point (**SP**) design based on the RDF **g** becomes no longer available, as shown in Figure 1.11 (b). Due to the long RTN **g** effect, the **h** is significantly shifted compared with Figure 1.11 (a). As a result, the probability of the pdf for the **h** becomes much higher at the point of **SP** (x=6), compared with Figure 1.11 (a). This means that unacceptable amount of the bits can be **out of spec** and causes some trouble in the market, i.e., after the shipment, as shown in Fig. 1.11 (b) and Fig. 1.12.
Figure 1.12 shows the concept of happening of reliability trouble caused by increased bit-fails after shipment due to time dependent margin variations.

To avoid this kind of reliability trouble, we have to precisely predict how much the tail of the whole margin distribution is shifted or extended in 10-years after shipment. Its required accuracy level in this work is less than the pdf=10^{-12}, which corresponds to 1bit fail of 1000-chips of 1G-bit SRAM.

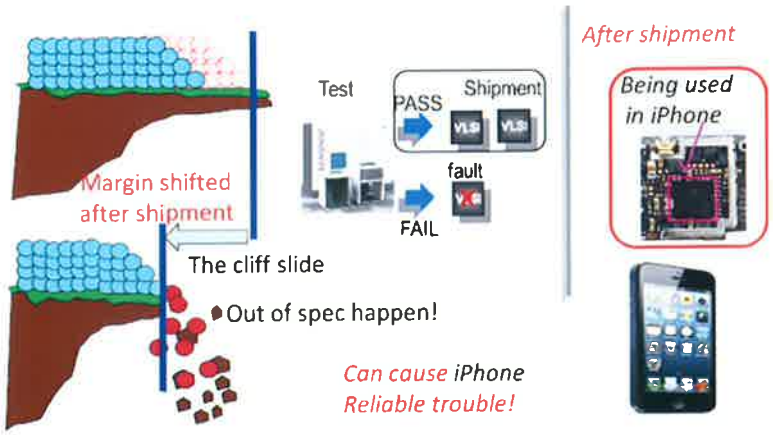


Figure 1.12: Concept of happening of reliability trouble caused by increased bit-fails after shipment due to time dependent margin variations.

1.2 Objectives of This Study

The subjects of this study are divided into two main categories: 1) approximation and decomposing for the complex distributions of the SRAM margin variations with EM

algorithm [8] with Gaussian Mixtures and 2) deconvolution algorithm [9] for reversing the RTN effects on the SRAM margin variations. The main purpose of use of those algorithms is to design a new VLSI's design style, as mentioned in Figure. 1.9.

Basic methodologies for the Expectation-Maximization algorithm and deconvolution algorithm are already developed, well known, and widely available to use by using the commercial product tools, such as MATLAB®-built-in functions. Main use applications for those are image processing and other signal processing. Main motivation for starting this study was the experiences that those algorithms could not be used for the application for advanced VLSI's design analyses. The potential reasons must come from the differences of the boundary conditions to be determined to solve the problem, compared with those for image processing.

Before showing how the conventional EM algorithm and deconvolution algorithm cannot work well for advanced VLSI's design analyses, the difference of the boundary conditions are explained first.

Figures 1.13(a) and 1.13(b) show an original gray scale of Lena image and one row 1-D image line intensity distribution of the Lena image, respectively. It is found that amplitude of intensity variations is within 2-orders range, as shown in Fig. 1.13(b).

On the other hand, the distribution of the target application of this study such as the RTN variation amplitude is spread in the range of 12-orders of magnitude, as shown in Fig. 1.13(c). This remarkable difference between the two applications leads to a completely different results. The details are described in sections 1.2.1 and 1.2.2 for the EM algorithm and deconvolution algorithms, respectively.

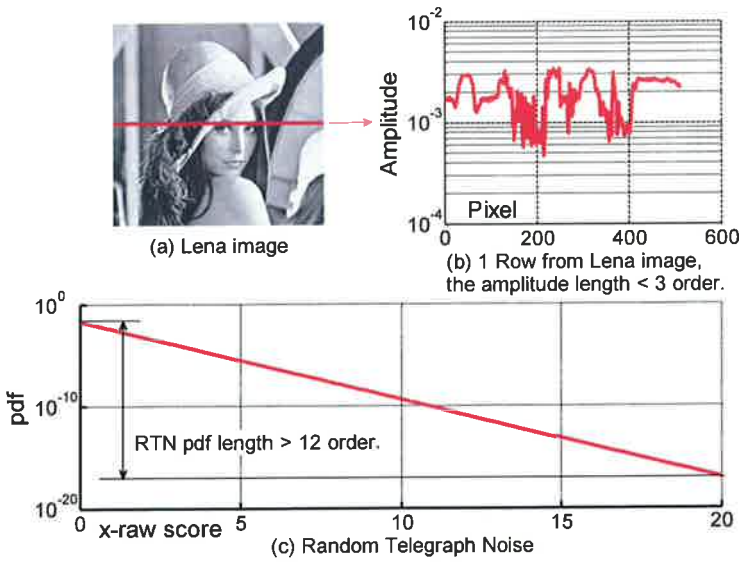


Figure 1.13: Comparison of an image signal and RTN distribution.

1.2.1 EM Algorithm for Approximation with Gaussian Mixtures

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

Figure 1.14 shows the comparisons of the GMM approximation results between (a) the 1D-Lena image data and (b) RTN data for this study application. It is found that the GMM fitting for the image data works well across the all range but doesn't fit at all for the attention tail region of the RTN effects variation data, as shown in Figure 1.14(b).

The main reasons for using GMM are: 1) it is easy to calculate based on error function based on cumulative density function (cdf) and 2) whole distributions can be divided into each element distribution based on possible impact factor and it obeys to Gaussian distribution. However, the conventional fitting algorithm tends to prioritize non-tail region because of giving much higher impact on increasing likelihood than that for tail region shown as Figure 1.14(b). Since the attention region for analyzing the fail-bit counts is in the tail region (at probability of 10^{-12}), it can be concluded that a new and modified idea is needed for the application of this study.

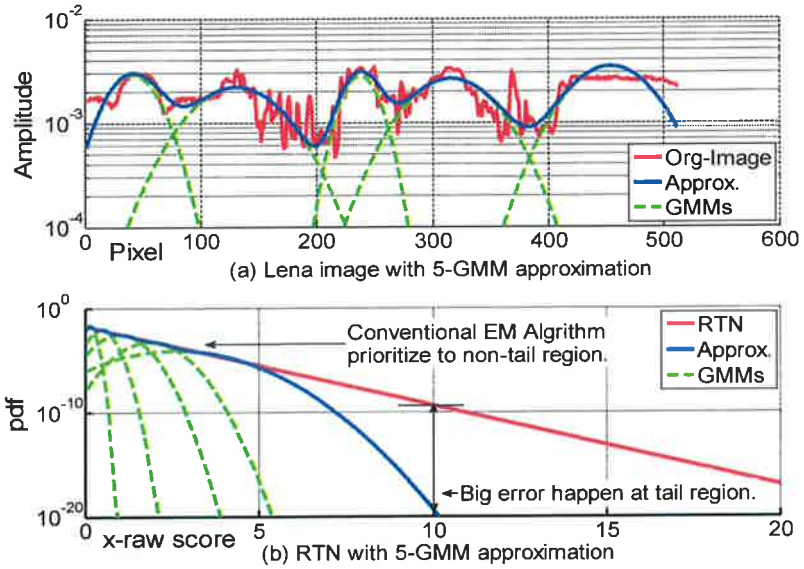


Figure 1.14: EM-Algorithm with 5 GMMs approximation of (a) image signal and (b) RTN distribution.

1.2.2 Deconvolution

Figure 1.15 shows the comparisons of the MATLAB®-built-in Richardson-Lucy deconvolution results between (a) the 1D-Lena image data and (b) RTN data for this study

application. It is found that the Richardson-Lucy deconvolution [10,11,12] results for the image data regenerate well across the all range but do exhibit the ringing for the attention tail region of the RTN effects variation data, as shown in Figure 1.15(b).

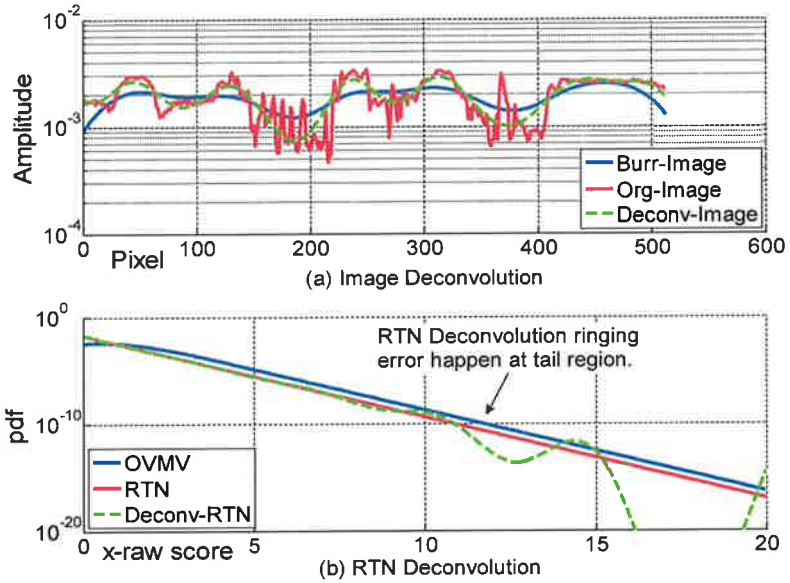


Figure 1.15: Richardson-Lucy deconvolution of (a) image signal and (b) RTN distribution.

The main reason for the deconvolution is the need for a reverse engineering to know what RTN **g** distributions effects the attention tail region of overall margin variations **h**=f⊗g. However, if the ringing happens like the case for MATLAB®-built-in Richardson-Lucy deconvolution, as shown in Figure 1.14(b), intolerable deconvolution error has to be expected. Thus, it was concluded that the MATLAB®-built-in Richardson-Lucy deconvolution cannot be used for this purpose. Thus, a new and modified idea is needed for the target application in this study, allowing to: 1) substantially circumvent the abnormal V-shaped ringing errors by eliminating the need of the inverse operation; 2) increase the deconvolution accuracy in the tail region where its pdf is less than 10^{-12}. Since the SRAM fail probability is extremely rare-event level (pdf is less than10^{-12}), the degree of precision for the tail distribution gives a big impact on the accuracy of the fail-bit count (FBC) prediction, and 3) guarantee the good enough deconvolution precision even if the RTN distribution comprises the complex gamma mixtures with the multiple convex and concave folding points.

1.2.3 Summary of Conventional Algorithm and Its Inherent Issues

Table 1.2 shows the summary of the conventional algorithm and its inherent issues for the EM algorithm and the deconvolution.

Table 1.2 Summary of conventional algorithm and its inherent issues.

Item	Conventional algorithm	Conventional inherent issues
EM algorithm for approximation with Gaussian mixtures	EM algorithm for approximation with Gaussian mixtures	Larger error in tail region i.e., less-populated region
Deconvolution	1) Algebraic	Ringing due to division by zero, high-frequency noise
	2) Algebraic + Digital filter	Inadequate ringing control Still ringing issues remain
	3) Iterative forward based i.e., Richardson-Lucy	Low frequency ringing noise due to 1) descent method used in Richardson-Lucy algorithm 2) Phase mismatch

1.2.4 Summary of Approaches in This Study to Solve the Conventional Issue

Table 1.3 shows the summary of the taken approaches in this study to solve the conventional issues.

Table 1.3 Summary of approaches in this study to solve the conventional issues.

Item	Conventional inherent issues	Solution given by this study
EM algorithm for approximation with Gaussian mixtures	Larger error in tail region i.e., less-populated region	Population-independent segmented approximation 1) Adaptive segmentation 2) Copy and paste fashion 3) Look up table based
Deconvolution	1) Algebraic 2) +Digital filter 3) Iterative forward based i.e., Richardson-Lucy	Iterative forward based 1) Segmentation 2) +Least square regression

As for the EM algorithm, the conventional fitting algorithm tends to prioritize non-tail region because of giving much higher impact on increasing likelihood than that for tail region, as shown in Figure 1.14(b). Since the attention region for analyzing the fail-bit

counts in the rare events is in the tail region (at probability of 10^{-12}). Thus, the following key ideas have been proposed: 1) adaptive segmentation to avoid a localized approximation, 2) copy and paste fashion with the scalable common kernel distribution and 3) look up table based determination of the kernel distribution.

As for the deconvolution algorithm, in this study, the three novel deconvolution methods for the SRAM-designs are proposed to eliminate the ringing issues: 1) segmentation to approximate with simple function so that the parameters for the function can be stable sought by the MATLAB®-built-in fmin-search function, 2) least square regression to reduce the number of parameters to be sought by the fmin-search function.

1.3 Organization of the Dissertation

The rest of the chapters are organized into the following eight chapters. Segmented EM algorithm is proposed to increase the accuracy of fitting a non-Gaussian long tail distribution with the Gaussian mixtures in Chapter-2. Table-lookup based algorithms are proposed to expand the versatility of type of distributions in Chapter-3. Potential applications of solving the forward problems/inverse problems, i.e., convolutions/non-blind and blind deconvolutions in VLSI design flows are introduced and discussed in Chapter-4. The proposed algorithms and its verification results are discussed and demonstrated in Chapter-5 following the verifications of the issues facing the conventional algorithms in Chapter-6. In Chapter-7, the proposed algorithms are further evolved to overcome some obstacles still facing the proposed ones when adopting the some unique distributions. Blind deconvolution algorithms and its verification results are proposed and demonstrated while comparing with the results for the non-blind cases in Chpter-8. To summarize this thesis, a brief summary for main achievements is given followed by describing the prospective challenges and future work in Chapter-9.

1.4 References

[1]. G. Moore, "VLSI: Some Fundamental Challenges," IEEE Spectrum, pp. 32, April 1979.

[2]. C. Shin, X. Sun, and TKJ Liu, "Study of Random-Dopant-Fluctuation (RDF) Effects for the Tri-gate Bulk MOSFET", IEEE Transactions on Electronic Devices, Vol.56, No.7, pp. 1538-1542, July 2009.

[3]. K. Takeuchi, T. Nagumo, K. Takeda, S. Asayama, S. Yokogawa, K. Imai, Y. Hayashi, "Direct Observation of RTN-induced SRAM Failure by Accelerated Testing and Its Application to Product Reliability Assessment " , in Digest of IEEE Symposium on VLSI Technology 2010, pp. 189-190, 2010.

[4]. K. Takeuchi, T. Nagumo and T. Hase "Comprehensive SRAM Design Methodology for RTN Reliability", Symposium on VLSI Technology, pp. 130-131, June 2011.

[5]. X. Wang, A.R. Broun, B. Cheng, A. Asenov, "RTS amplitude distribution in 20nm SOI FinFETs subject to Statistical Variability", Simulation of Semiconductor Processes and Devices SISPAD 2012, pp.296-299, 2012.

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

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

[8]. Moon, T.K, "The expectation-maximization algorithm" in Signal Processing Magazine, IEEE, Volume: 13, Issue: 6, pp. 47-60, 1996.

[9]. Q. Shi, O. Kanoun, "Application of Iterative Deconvolution for Wire Fault Location via Reflectometry", Digest of IEEE International Symposium on Instrumentation & Measurement, Sensor Network and Automation (IMSNA), pp. 102-106, 2012.

[10]. Fish D.A. A.M. Brinicombe, E.R. Pike, "Blind deconvolution by means of the Richardson–Lucy algorithm", Journal of the Optical Society of America A12 (1), pp.58–65, 1995.

[11]. White R. L. "Image Restoration Using the Damped Richardson-Lucy Method" The Restoration of HST Images and Spectra II Space Telescope Science Institute, pp.104-110, 1994.

[12]. F. Dell Acqua1, "A Modified Damped Richardson-Lucy Algorithm to Improve the Estimation of Fiber Orientations in Spherical Deconvolution", Proc. Intl. Soc. Mag. Reson. Med. 16, pp. 1860, 2008.

Chapter 2

Segmented Gaussian Mixtures EM Algorithm for Fitting Non-Gaussian Distributions

In this chapter, we propose a fitting method to approximate the mixtures of various sloped-tail Gamma distributions characterizing the RTN by an adaptive segmentation 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.

2.1 Introduction

The approximation-error of the tails of 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 Figure 2.1. Since the increasing paces of variation-amplitude ΔV_{th} are differently dependent on the MOSFET channel-size (LW) like the below expressions of (2.1) and (2.2), the ΔV_{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 Figure 2.1(a).

$$\Delta \mathrm{Vth\,(RDF)} \propto \mathrm{AVt\,(RDF)} \,/\sqrt{\mathrm{LW}} \qquad \text{--------- (2.1)}$$

$$\Delta \mathrm{Vth\,(RTN)} \propto \mathrm{AVt\,(RTN)} \,/\, \mathrm{LW} \qquad \text{--------- (2.2)}$$

Where, AVt (RDF) and AVt (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 Figure 2.1. (a). According to the references [1]-[5], there will come the time soon around 15nm scaled CMOS era.

The issues caused by the size scaling will be discussed in this chapter. Figure 2.1 (b) illustrates the probability density functions for RDF, RTN1(40nm), RTN2(<16nm) and RTN3 (<7nm), and its convolution results, respectively.

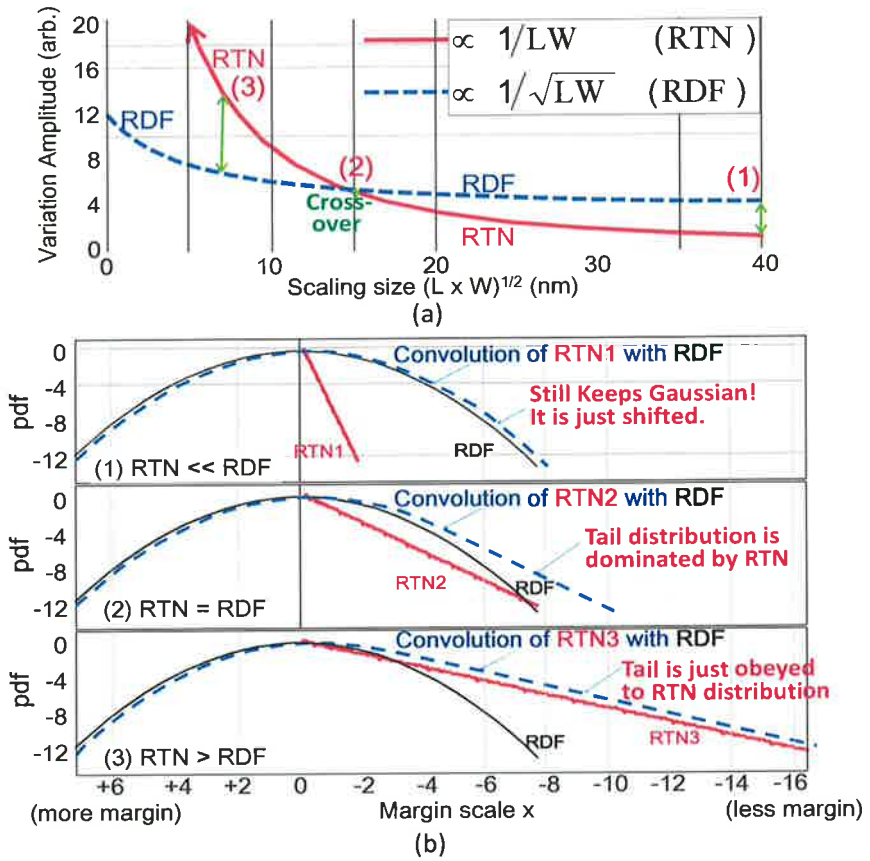


Figure 2.1: (a) Trend of variation amplitude of RTN and RDF (b) Comparisons of convolution results between 3 cases for 40nm, 16nm, and 7nm.

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. 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 Figure 2.1. (b).

In order to solve the above issues, this study proposes, 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 β=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 study 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 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 fail-bit counts before and after screening at the stages of both circuit design and screening test.

2.2 Discussions on the Conventional Models

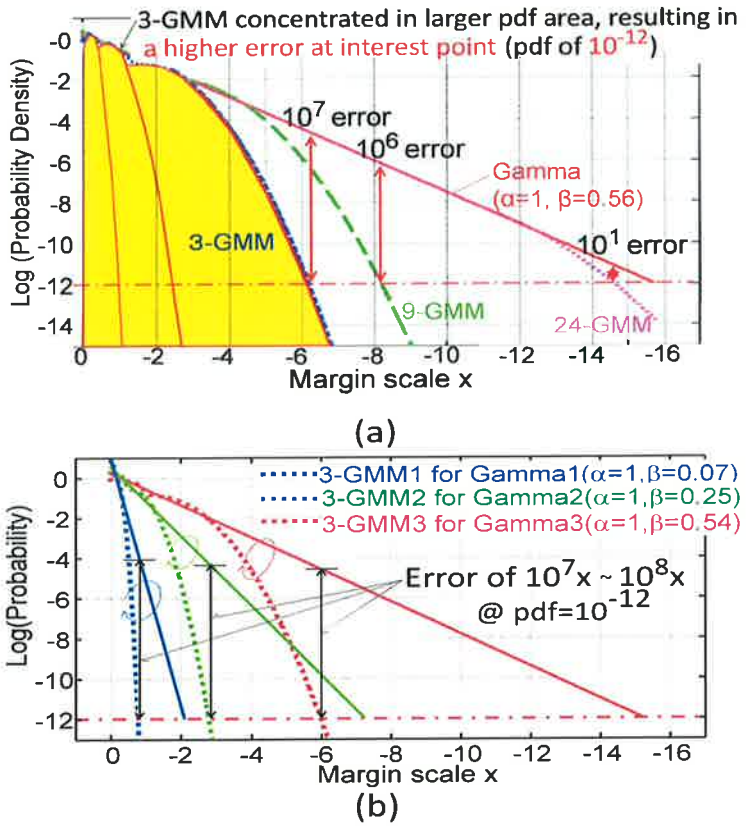


Figure 2.2: (a) Approximation error comparisons between 3, 9, 24-GMMs (b) Error dependency of 3 types of Gamma distributions of β=0.07, 0.25 and 0.54.

Expectation-maximization (EM) algorithm [6], which is an iterative procedure that maximizes the likelihood of GMM, is well known as easy and convenient means to approximate 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 Figure 2.2. Since the interest region for analyzing the fail-bit counts of the rare-events is in the tail region (at probability of 10^{-12}), the EM algorithm for this application leads to a significant fail-bit count error of orders of 10^{7}, as shown in Figure. 2.2. 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 Figure 2.2. In almost all fail-bit analyses, the distribution of interest only matters in the tail-region of probability of orders of 10^{-12} [1]- [3]. Thus, this is a crucial challenge that should be addressed until the time comes for the rare-event SRAM yield predictions.

2.3 Proposed Statistical Approximation Model for RTN Gamma Distribution

In order to solve these crucial issues, this study develops 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 Figure 2.3(a) and (b), respectively.

a) Adaptive segmentation

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

(1) 1st-step is to do approximation by 3-GMM between X0 and Xn. Then 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. Then 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. Then 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 Figure 2.3(a).

b) Copy and paste fashion

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. Then, 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 1st 3-GMM.

(2) 2nd-step is to get the weight (w_1) of the 2nd 3-GMM. Then, 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= $(w_0 - w_1)/ \Delta X$

(3) 3rd-step is to do the same thing as (2), as shown in Figure 2.3(b). This flow can be repeated until Xm>Xn.

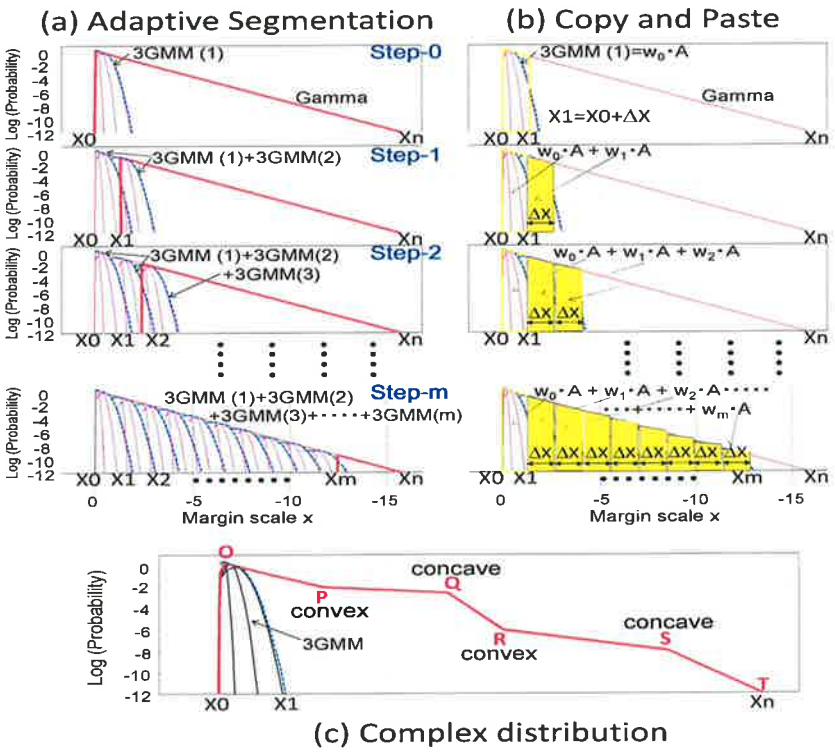


Figure2.3: Concepts of the proposed approximation algorithm. (a) adaptive segmenttaion (b) copy and paste fashion (c) example of complex distributions.

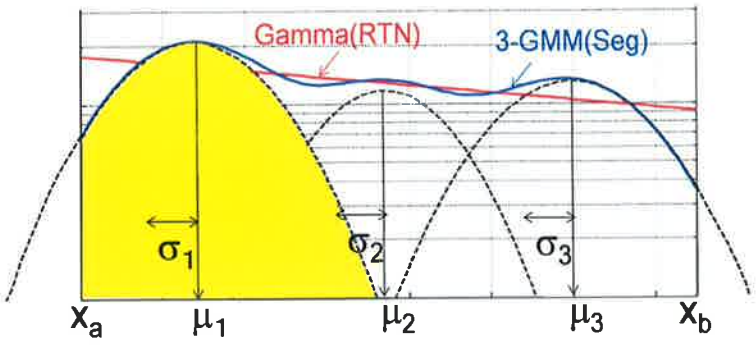
Table 2.1: Comaprisons of EM-iterations.

	Segmentation (proposed)	Conventional (w/o segmentation)
Gamma1	3	338
Gamma2	6	340
Gamma3	6	340

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 Figure 2.3(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 2.6, 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 2.1. This also allows us to use only Gaussian distributions when doing convolution of Gaussian-RDF and Non-Gaussian-RTN distributions.

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 caluculate the fail-bit error counts of the segmentation of (x_a-x_b) is shown in Figure 2.4.



S_1=normcdf(x_b,μ_1,σ_1)-normcdf(x_a,μ_1,σ_1)
S_2=normcdf(x_b,μ_2,σ_2)-normcdf(x_a,μ_2,σ_2)
S_3=normcdf(x_b,μ_3,σ_3)-normcdf(x_a,μ_3,σ_3)

Figure 2.4: Error bit counts of the segmentation can be given by normcdf of three GMMs.

2.4 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:

(σ=1, μ=0) for Gaussian, (α=1, β=0.56) for Gamma. The relationship between the two distributions and its convolution are shown in Figure 2.1(b) and 2) comparisons of the following 6 approximation-models of Gamma distribution (α=1, β=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.

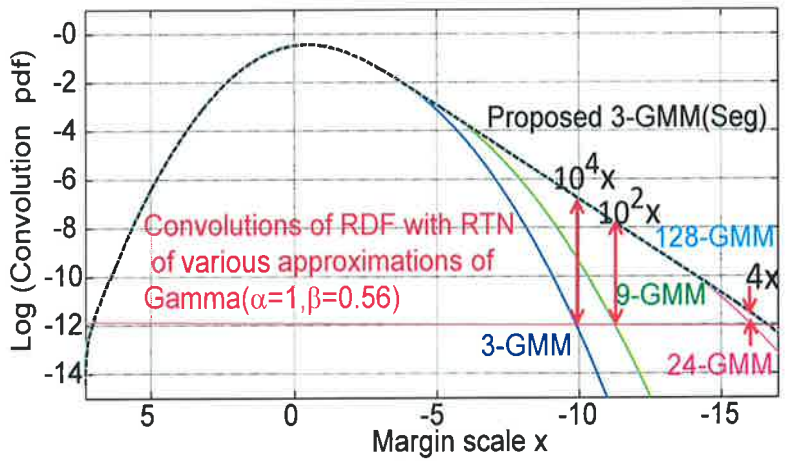


Figure 2.5: Comparisons of the convolution results between the cases of 3, 9, 24, 128-GMMs and the proposed.

Figure 2.5 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 7-orders and 4-orders of magnitude at the fail probability of 10^{12}, as shown in Figure 2.5. It is worth mentioning that 3-GMM segmentation scheme provides a better approximation than the case of 24-GMM, as shown in Figure 2.5 In order to characterize the error of each convolution result, the "golden", which is given by the numerical calculation of convolution of Gaussian and Gamma distributions, is used as a reference. The numbers of fail-bit count errors for each approximation model are compared, as shown in Figure 2.6. The number of fail-bit count error is defined as the difference in the cdf value between the "golden" and that for each model.

It is worth mentioning that the relationship of "which is better" is dependent on the x-scale, as shown in Figure 2.6. For example, 128-GMM is the best in X= (-6~-12) in which the fail-probability is larger than 10^{-8} (shown in Figure 2.6). In contrast, the proposed one can reduce the error best in X= (-12~-16) in which the fail-probability is smaller than 10^{-8} where there is the most interest point for the GB designs.

When discussing the GB designs for volume production, the expected yield-loss should be predicted. This study assumed that the target fail-probability is 10^{-12} to realize 99.9%

yield of 1Gbit memory chip, which is a quite realistic target. Thus, our most interested point of x is around x=-16, where fail-probability is around 10^{-12}, as shown by Figure 2.6. Our proposed 3-GMM segmentation method can provide the best approximation compared with others.

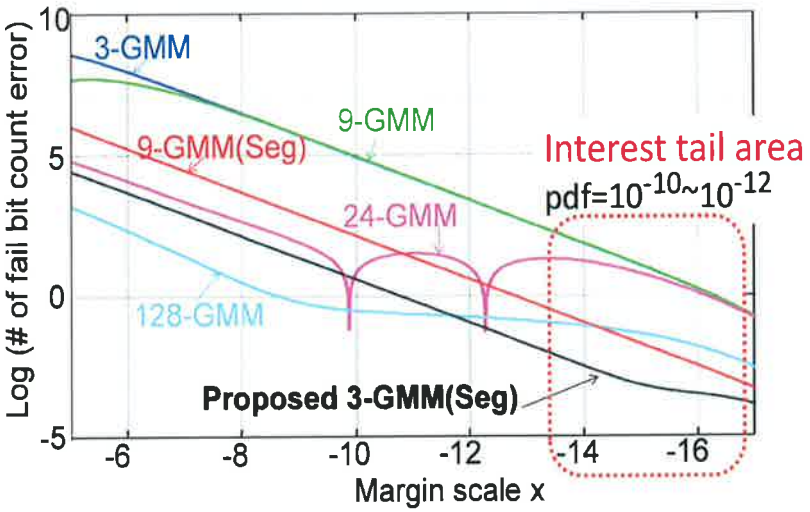


Figure 2.6: Comparisons of fail-bit count errors between 3, 9, 24, and 128-GMMs without segmentation and 3 and 9-GMM with segmentation.

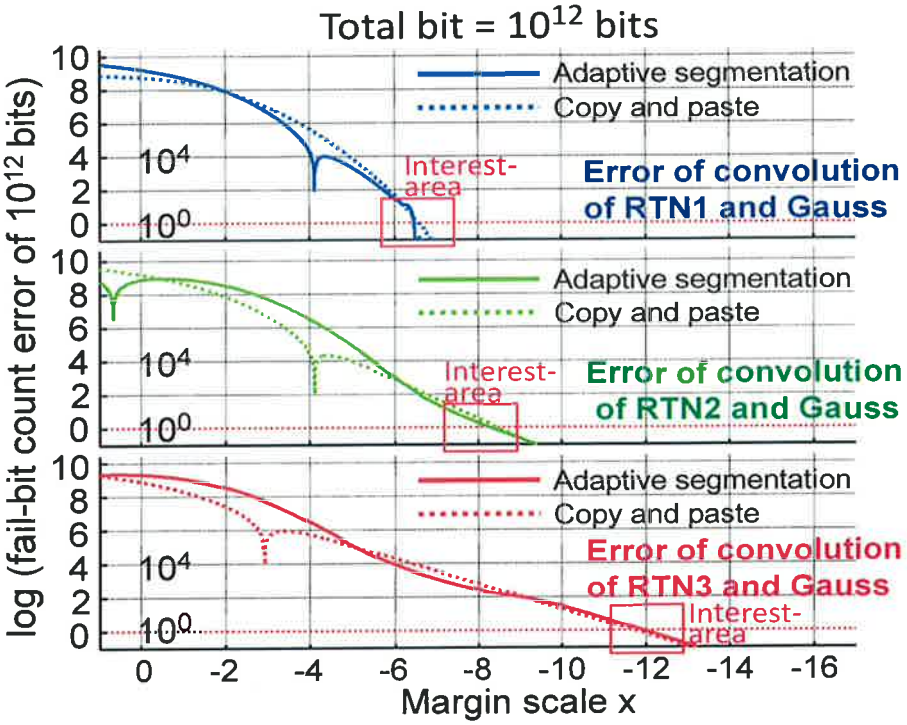


Figure 2.7: Comparisons of fail-bit count errors between the adaptive segmentation and copy and paste fashion for the 3-cases of RTN1, RTN2, and RTN3.

As mentioned earlier, 9-GMM segmentation is worth than 3-GMM segmentation in the wide range of x=-6~-16 because the variation of probability becomes larger and the density of GMM in area of lower probability becomes much smaller as shown in Figure 2.6. Regarding the cases without using segmentation, the number of errors is 1~4 orders of magnitude larger than that for the proposed 3-GMM segmentation scheme.

Figure 2.7 shows the comparisons of the numbers of fail-bit count errors between the adaptive segmentation and copy and paste fashion proposed in this chapter. The 3-cases of the errors of the convolution of Gauss with RTN1, RTN2, and RTN3 are shown, respectively. Although the small difference in terms of the fail-bit counts can be seen in the non-interest area, it is found that the both ideas of the proposed adaptive segmentation and copy and paste fashion can provide the small enough accuracy in the interest area (rare-event), compared with the conventional means, as shown in Figures 2.6-2.7

2.5 Application to More Complex Distributions

According to the reference [7], 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 Figure 2.8. 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 Figure 2.8. 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 Figure 2.8.

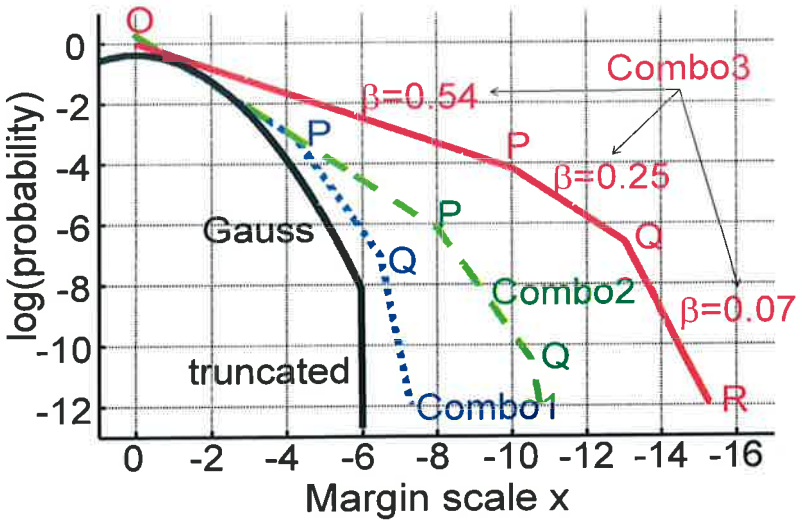


Figure 2.8: Distributions of Gauss (RDF) and combination shaped gamma distributions of Combo1, Combo2, and Combo3.

The proposed ideas of "adaptive segmentation" and "copy and paste" fashion can be applied 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).

Figure 2.9 shows the comparisons of approximation-errors for fitting to Combo1, Combo2, and Combo3 between the cases of (a) using the convnetional 3-GMM model and (b) using the proposed segmentation models. As can be seen in the Figure 2.9(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 Figure 2.9(b). Unlike the case of Figure 2.9(a), it can be seen that the fitting curves and its target lines in Figure 2.9(b) are perfectly overlapped. Thanks to the segmentation manner, the same concepts can be adaptively applied to the different sloped-tail distributions. Thus, these ideas can be applied to the various sloped-distributions even if they are combined like the given examples in Figure 2.8.

Since the both ideas of "adaptive segmentation" and "copy and paste" fashion can be applied 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 Figure 2.10.

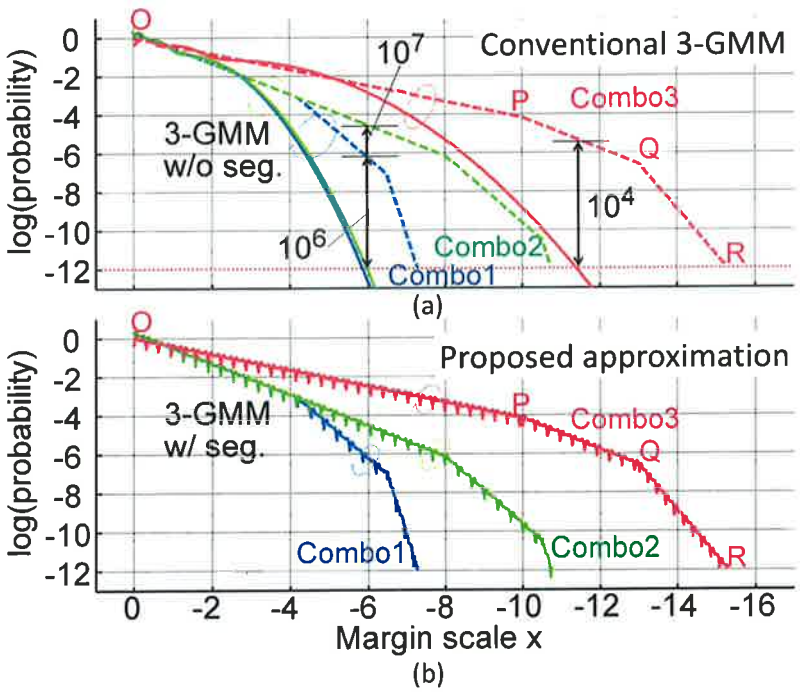


Figure 2.9: Comparisons of errors for Combo1, Combo2, and Combo3 between (a) with the convnetional 3-GMM and (b) with the proposed segmentation.

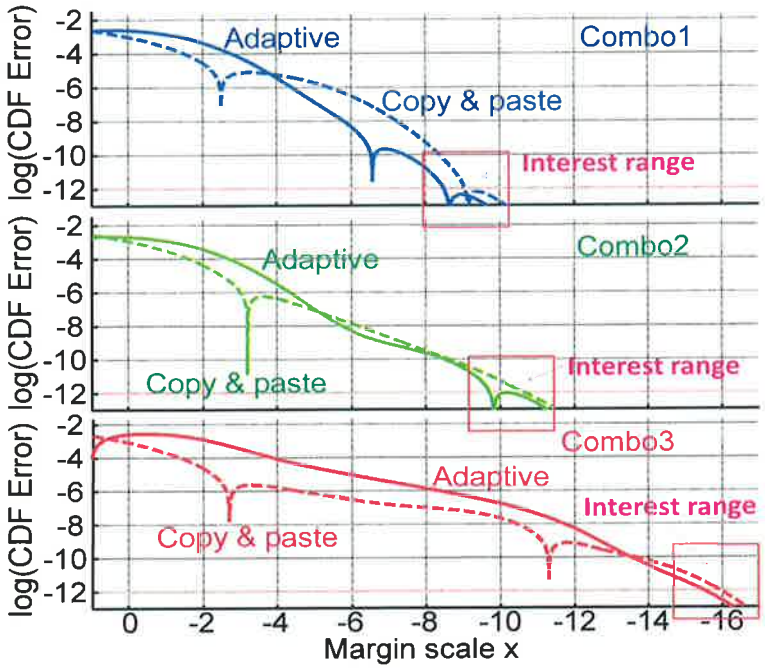


Figure 2.10: Comparisons of cdf error for Combo1, Combo2, and Combo3 between the adaptive segmentation and copy and paste.

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 Figure 2.10.

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.

2.6 Conclusions

This chapter discusses, 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 cannot be used 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 but follows to the mixtures of various-sloped Gamma distributions.

To address the above issues, the algorithms have been proposed, which can approximate the tails of RTN distributions with a convenient and simple GMM. This allows the fitting model to be applied to 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.

This study has 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.

2.7 References

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

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

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

[4]. K. Takeuchi, T. Nagumo and T. Hase "Comprehensive SRAM Design Methodology for RTN Reliability", Symposium on VLSI Technology, pp. 130-131, June 2011.

[5]. K. Takeuchi, T. Nagumo, K. Takeda, S. Asayama, S. Yokogawa, K. Imai, Y. Hayashi, "Direct Observation of RTN-induced SRAM Failure by Accelerated Testing and Its Application to Product Reliability Assessment " , in Digest of IEEE Symposium on VLSI Technology 2010, pp. 189-190, 2010.

[6]. Moon, T.K, "The expectation-maximization algorithm" in Signal Processing Magazine, IEEE, Volume: 13, Issue: 6, pp. 47-60, 1996.

[7]. X. Wang, A.R. Broun, B. Cheng, A. Asenov, "RTS amplitude distribution in 20nm SOI FinFETs subject to Statistical Variability", Simulation of Semiconductor Processes and Devices SISPAD 2012, pp.296-299, 2012.

Chapter 3

SRAM Screening Test Design and Lookup Table Based Fitting Algorithm for Complex RTN

In this chapter, how the challenges facing the guard-band designs should be addressed, including the margin assist-circuits scheme for the screening-test in the coming process generations. 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 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 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.

A lookup table (LUT) based fitting is new proposed method to approximate the mixtures of various sloped-Gamma tail distributions by an adaptive segmentation 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.

3.1 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 (V_{DD}) 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.

Because the required GB voltages will be no longer small fraction of the whole margins, as shown in Figure 3.1, the conventional GB design criteria with the screening test won't be effective any more for avoiding the out of spec after the screening.

The GB(TD) in Figure 3.1 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 Figures 3.1 and 3.2. 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 Figure 3.2. 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.

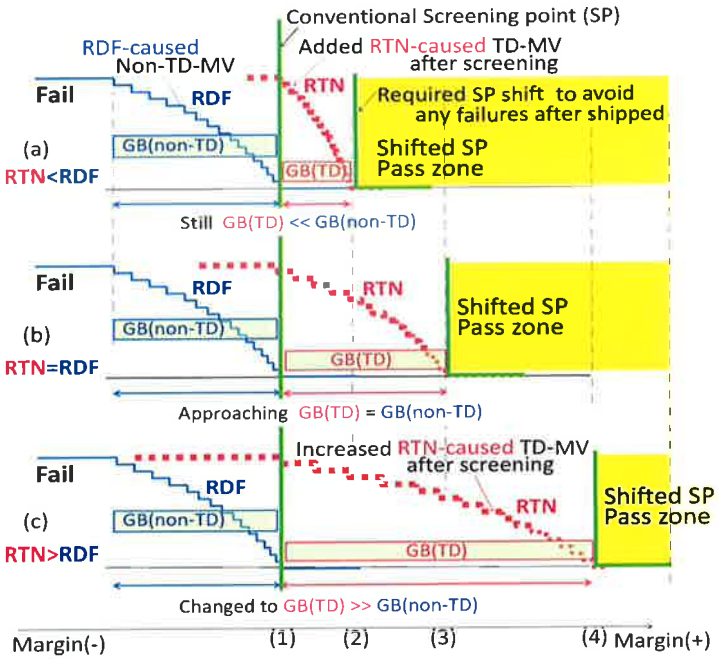


Figure 3.1: Change of the GB width ratio for non-TD to TD variations caused by RDF and RTN: (a) RTN<RDF, (b) RTN=RDF, (c) RTN>RDF.

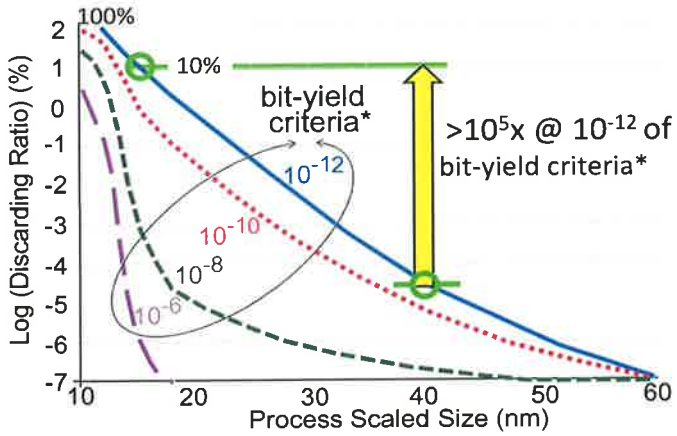


Figure 3.2: Increased chip discarding ratio for 15nm can be a 10^5 x larger than that for 40nm.

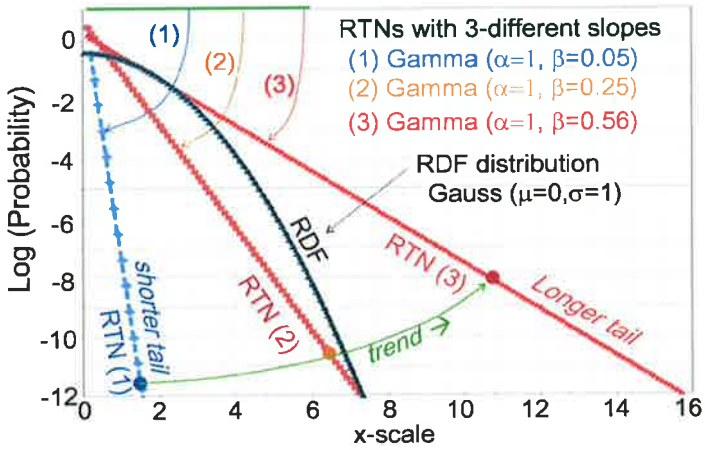


Figure 3.3: Relationships between the tail distributions for RDF (Gauss) and three RTN1, RTN2, and RTN3.

In order to discuss the impacts of rapidly increased the TD-MV on the GB designs, the three cases of the ΔV_{th} ratios of RTN/RDF: (1/4, 1/1, 4/1) are assumed in this discussion, as shown in Figure 3.3.

Marked (1), (2), and (3) in Figure 3.4 represent for the three cases of the relationship of the ratios of RTN/RDF, respectively. Three different RDF distributions, referred to as RDF1, RDF2, and RDF3 are assumed. 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 Figure 3.4.

In this study, 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.

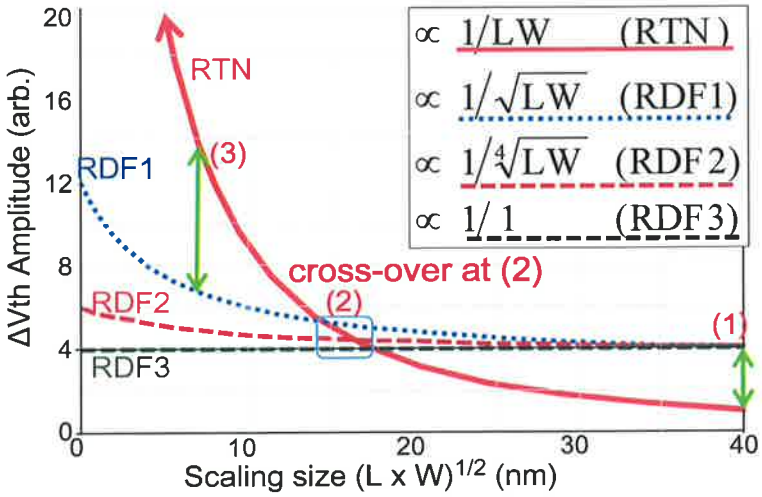


Figure 3.4: Trends of ΔV_{th} amplitude of RTN and RDF. The increasing pace of RTN is assumed as proportional to 1/LW.

3.2 Discussions on Issues of Guard Band Designs

The effects of long tail distributions on the shifted screening point (SP) are shown in Figures. 3.5-3.7. The different RTN amplitudes of RTN1, RTN2, and RTN3 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 Figures. 3.5-3.7. 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 Figures 3.5-3.7, 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 Figures 3.5-3.7, the Δx for RTN1, RTN2, and RTN3 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 RTN1), the distribution of the convolution results has a folding point like Q shown in Figure 3.5. In contrast, the

convolution for RTN2 and RTN3 is not likely to have any 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.

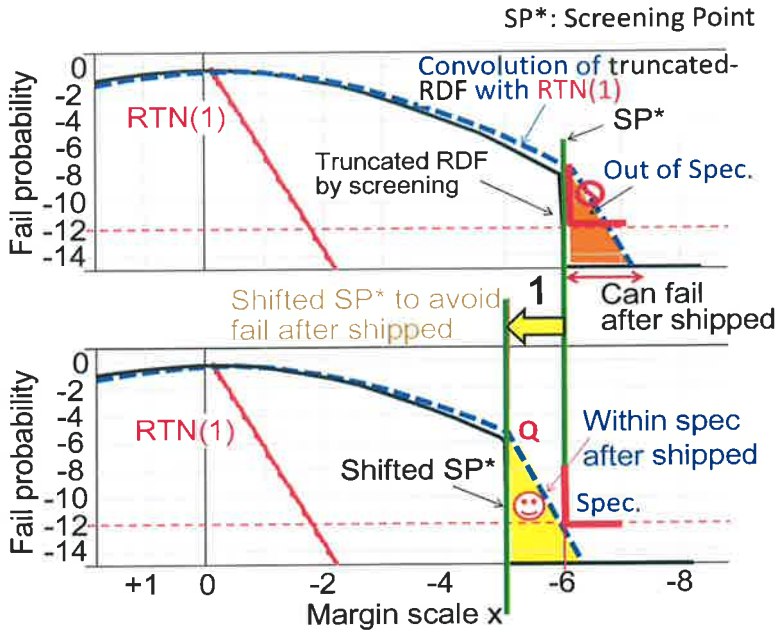


Figure 3.5: RTN1 impacts on the tails after the screening. To avoid any fail after shipped, the screening point has to be shifted by 1 of x.

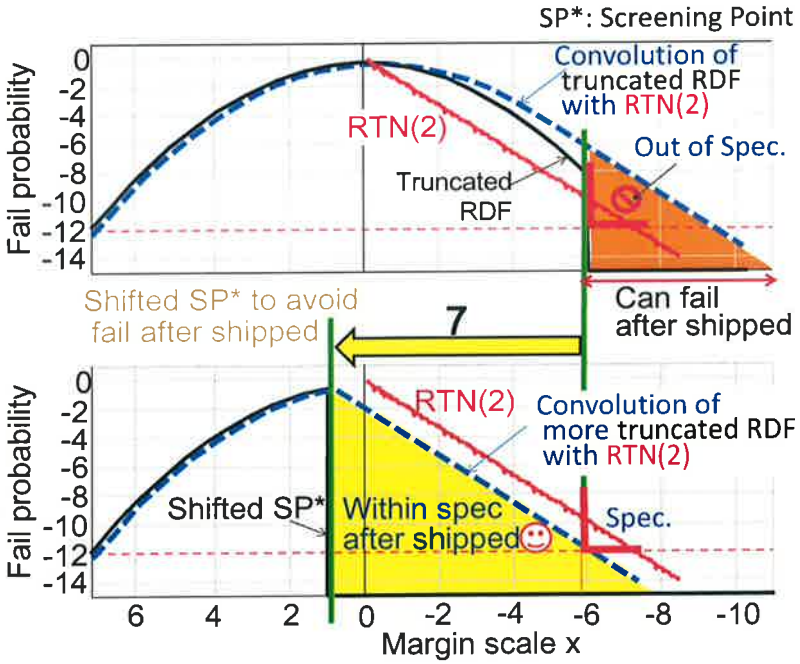


Figure 3.6: RTN2 impacts on the tails after the screening. To avoid any fail after shipped, screening point has to be shifted by 7 of x.

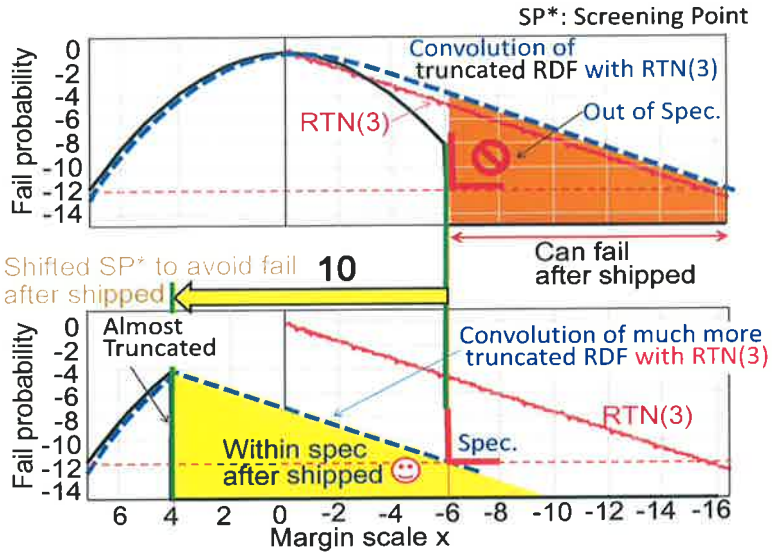


Figure 3.7: RTN3 impacts on the tails after the screening. To avoid any fail after shipped, screening point has to be shifted by 10 of x.

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.

The effects of the excessive chip-discarding yield-loss made by the error of the RTN approximation are shown in Figure 3.8. 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 study, as shown in Figure 3.8.

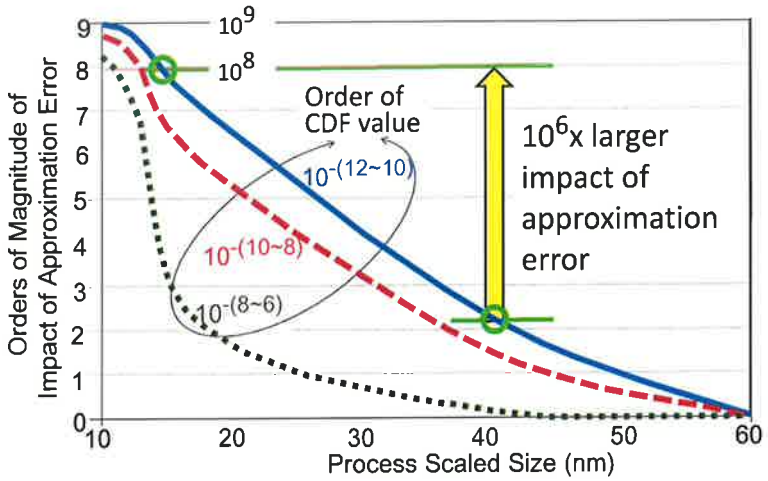


Figure 3.8: Increased impact of approximation error on the trouble of the excessive under-estimation/over- estimation of the yield.

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 Figure 3.13. As show 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.

3.3 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).

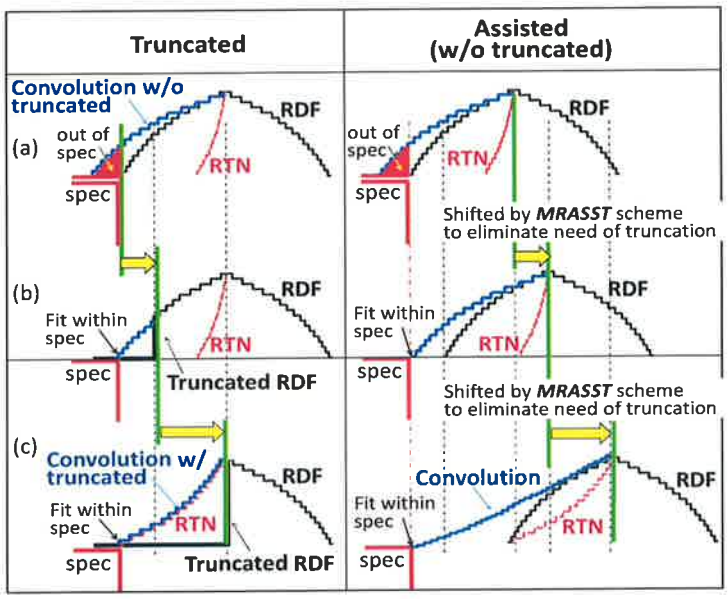


Figure 3.9: Increased impact of approximation error on the trouble of the excessive under-estimation/over- estimation of the yield.

As discussed in Figures. 5-7, the yield-loss will become prominent by the screening test unless adequately treated at the GB designs. To address this challenge, will need to design the margin assist circuits (MRASST) for the RTN-caused variations as shown in Figure 3.9 (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.

3.4 Challenges 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 Figure 3.1. 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 Figure 3.10.

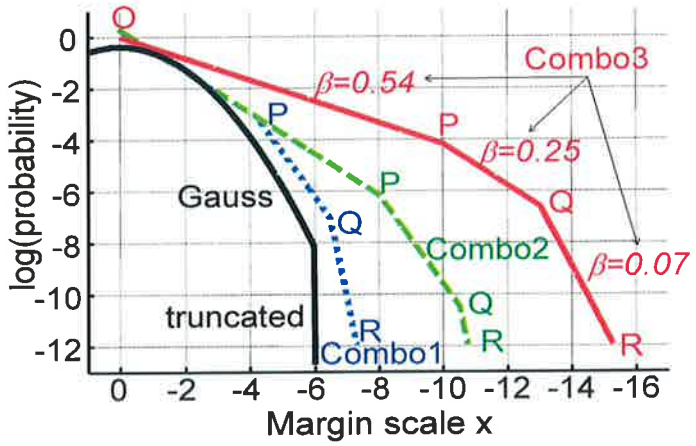


Figure 3.10: Distributions of Gauss (RDF) and combination shaped gamma distributions of Combo1, Combo2, and Combo3.

Figure 3.11 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.

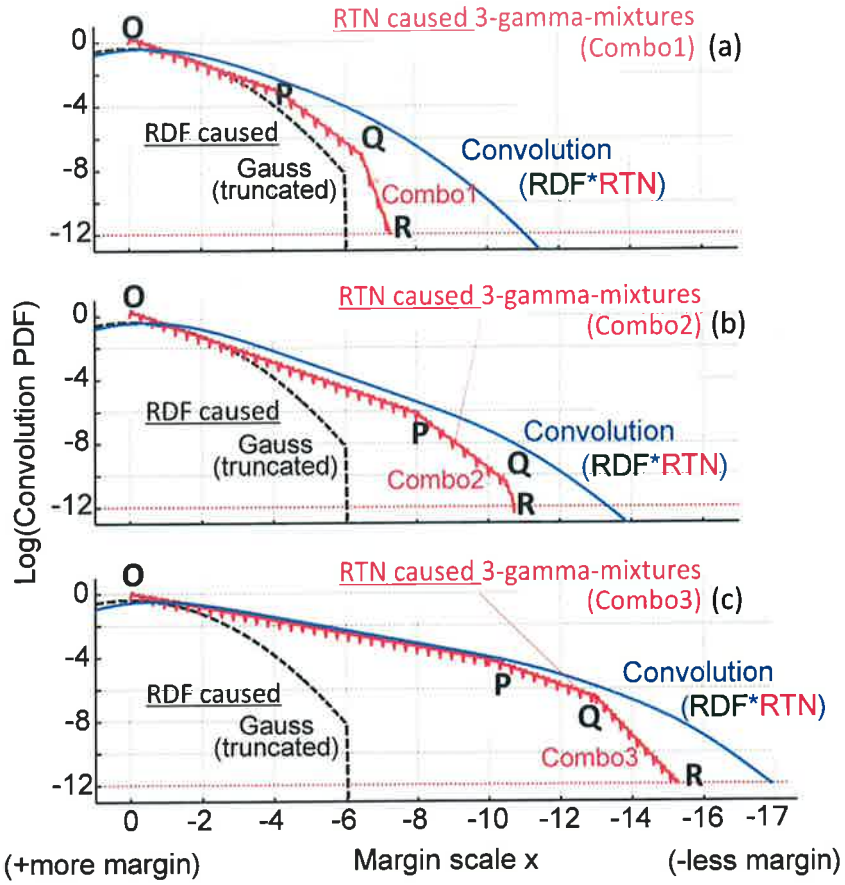


Figure 3.11: Comparisons of the convolution results between the truncated RDF and RTN (a) combo1, (b) combo2, and (c) combo3.

3.5 Proposed LUT Based Statistical Approximation Model for RTN Gamma Distribution

In order to solve these crucial issues, this study develops 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 approximations with lookup table describe as bellow.

a) Adaptive segmentation

Algorithm of the adaptive segmentation was described in section 2.4

b) Copy and paste fashion

Algorithm of the copy and paste fashion was described in section 2.4

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 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

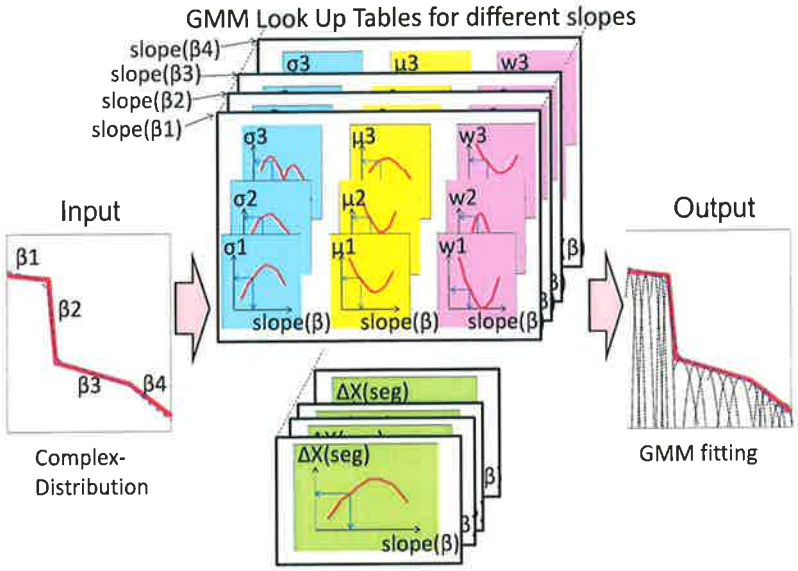


Figure 3.12: Concept of the look up table for the different sloped- gamma distributions of shape parameter β.

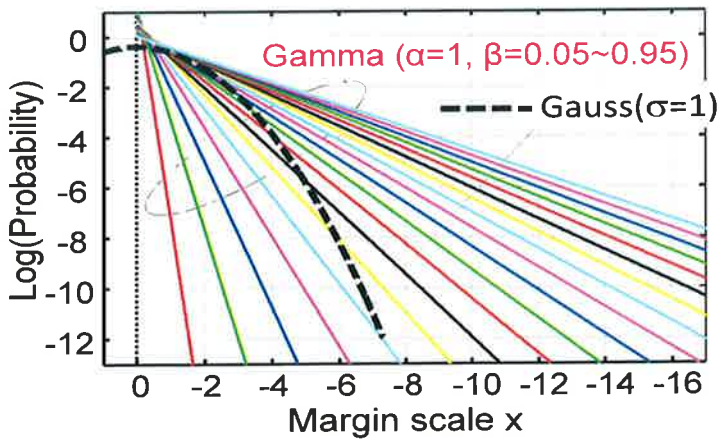


Figure 3.13: Various sloped RTN distributions compared with RDF distribution.

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 study 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 all parameters required to regenerate the GMM comprising 3-Gaussians, as shown in Figure 3.12.

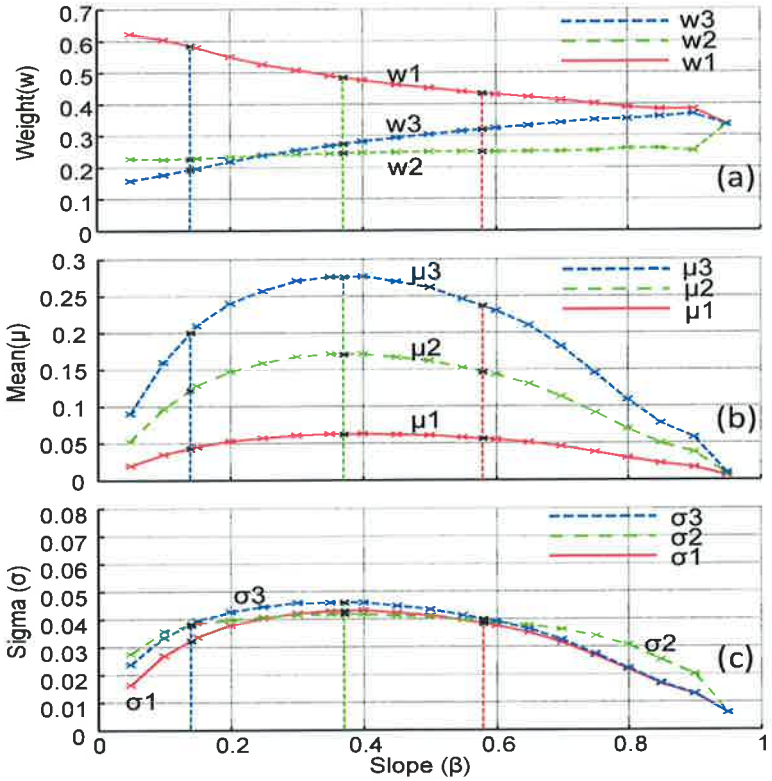


Figure 3.14: Slope dependency of parameters in look up table for three Gaussian mixture models.

This also outputs the best width of individual segmentation ΔX(seg) that can make the likelihood maximized. As a result, overall approximations with optimized segmentation width can be easily done without any time-consuming EM steps. 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 Figure 3.13.

Since the slope β dependencies of all parameters of GMM and the best width of individual segmentation ΔX(seg) have a simple and continuous relationship, as shown in Figure 3.14 Figure 3.15, the error cause by interpolation of LUT can be minimized as discussed in the following section 3.

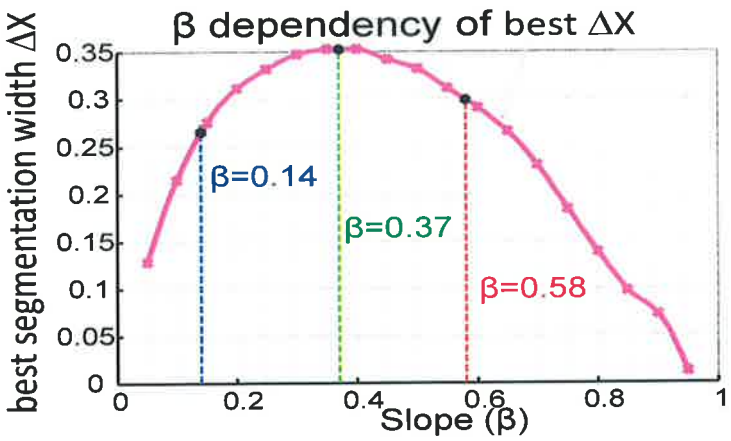


Figure 3.15: Slope dependency of parameters in look up table for three Gaussian mixture models.

Figure 3.16 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 Figure 3.16. Thus, if the slope β is input to the LUT, the ΔX is also given besides the parameter set for 3-GMMs (shown in Figure 3.14).

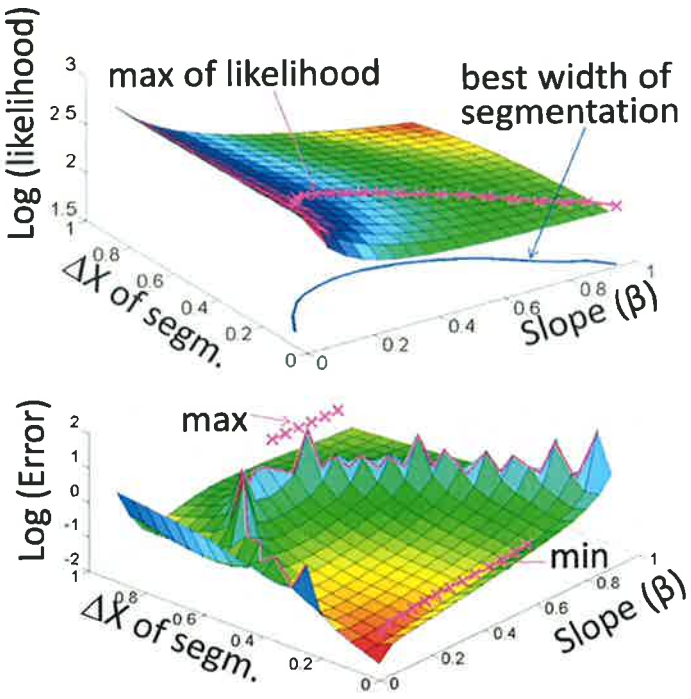


Figure 3.16: Likelihood and approximation error dependencies of the slope-β and the segmentation width-ΔX.

3.6 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 Figure 3.17.

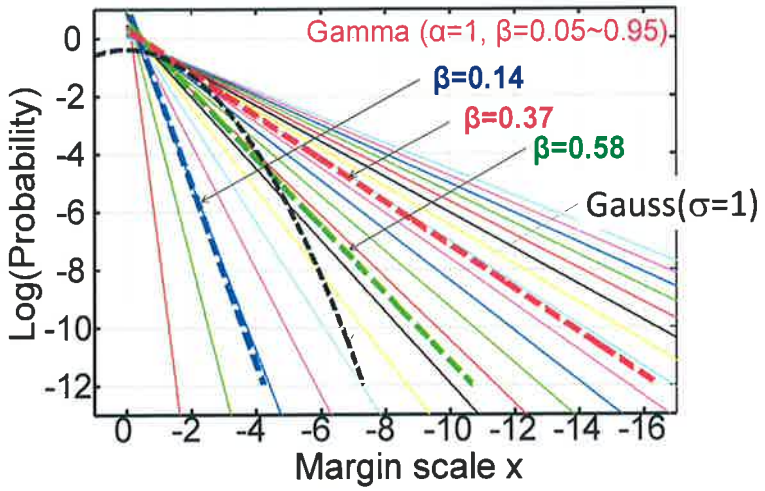


Figure 3.17: Likelihood and approximation error dependencies of the slope-β and the segmentation width-ΔX.

Figure 3.18 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.

Figure 3.19 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].

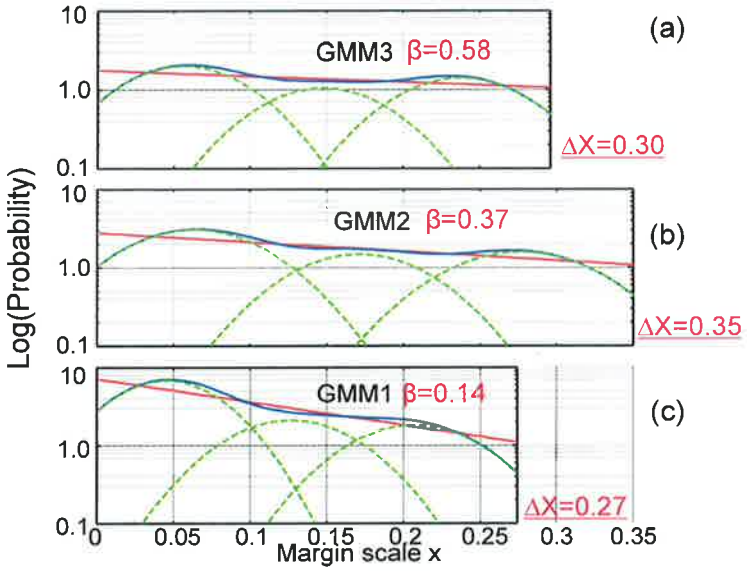


Figure 3.18: 3-GMMs in the different best segmentation for different sloped tails of (a) ΔX=0.3, β=0.58, (b) ΔX=0.35, β=0.37 (c) ΔX=0.27, β=0.14.

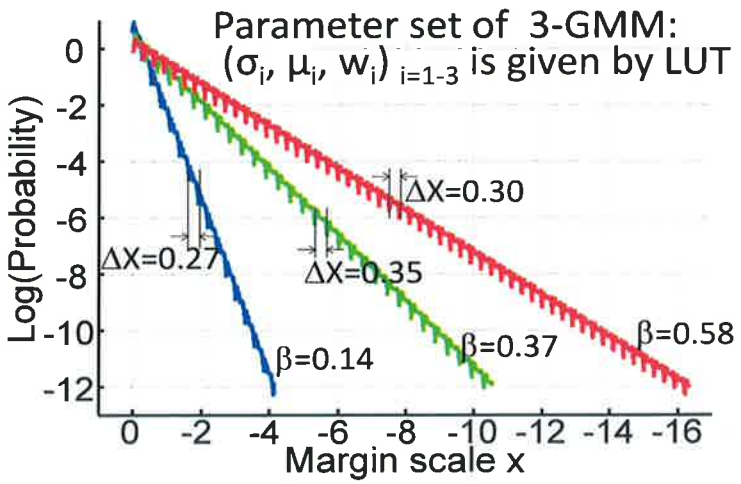


Figure 3.19: LUT based fitting of the different sloped tails of (a) ΔX=0.3, β=0.58, (b) ΔX=0.35, β=0.37, (c) ΔX=0.27, β=0.14.

Figure 3.20 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 interested region compared with the other two proposed schemes, the orders of error are compared in the interest region (cdf of 10^{-12}), as shown in Figure 3.20. 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.

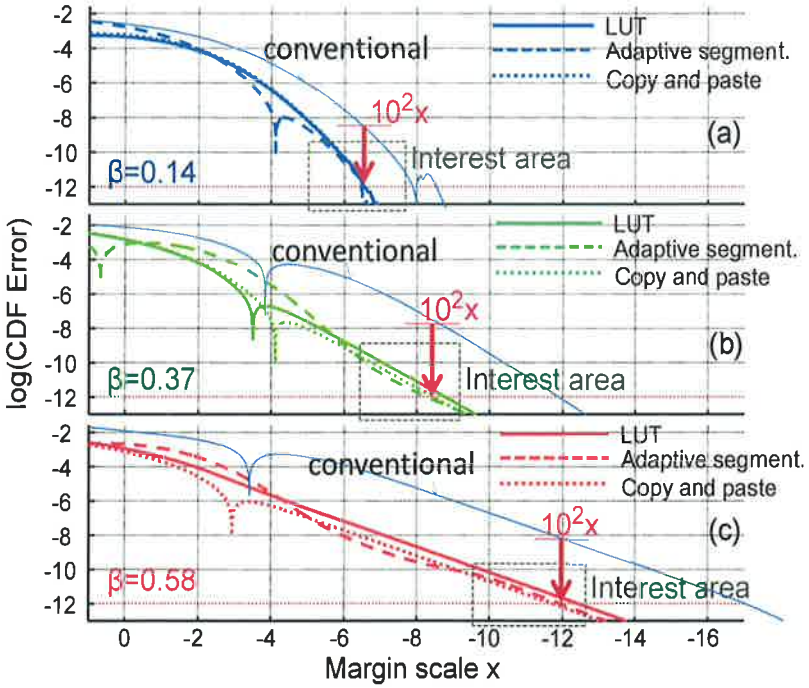


Figure 3.20: The cdf error of the convolution results between Gauss (σ=1) and 3-different Gammas of (a) β=0.14, (b) β=0.37 and (c) β=0.58.

3.7 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 Figure 3.21. 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 Figure 3.21.

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 Figure 3.21(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 Figure 3.21(b). Unlike the case of Figure 3.21(a), it can be seen that the fitting curves and its target lines in Figure 3.21(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 Figure 3.10.

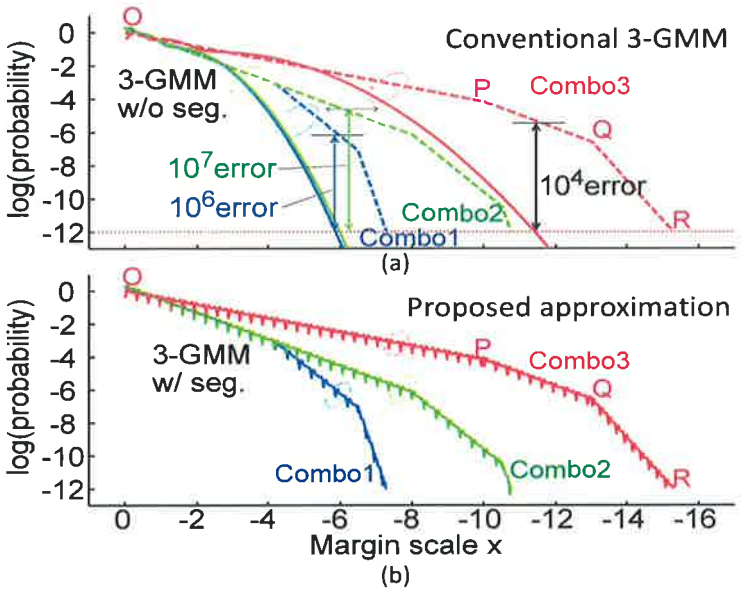


Figure 3.21: Comparisons of errors fitting to Combo1, Combo2, and Combo3 between (a) the convnetional 3-GMM and (b) the proposed segmentation.

Since the both ideas of "adaptive segmentation" and "copy and paste" fashion can be applied 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 Figure 3.22.

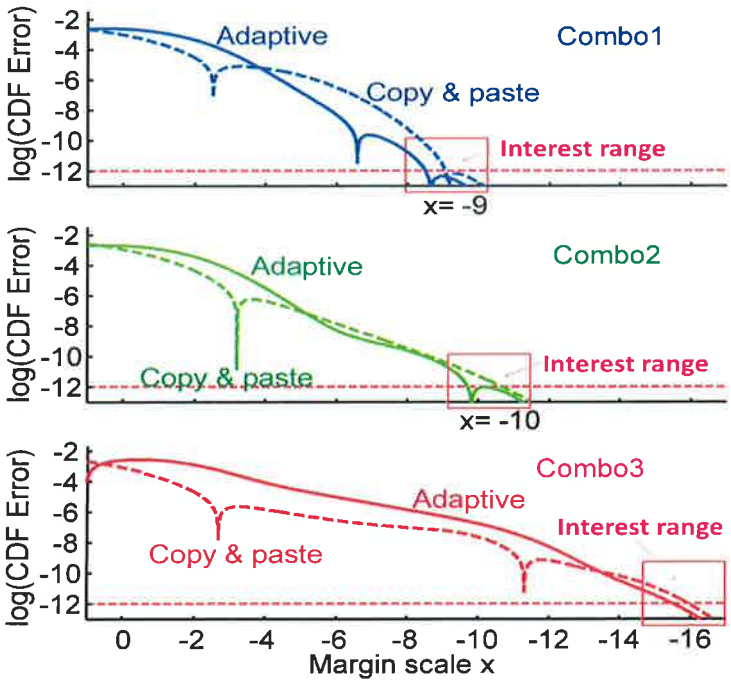


Figure 3.22: Comparisons of the cdf error of the convolution results for Combo1, Combo2, and Combo3 between "adaptive segmentation" and "copy and paste".

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 Figure 3.22. 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 interested 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 Figure 3.23(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 Figure 3.23(b).

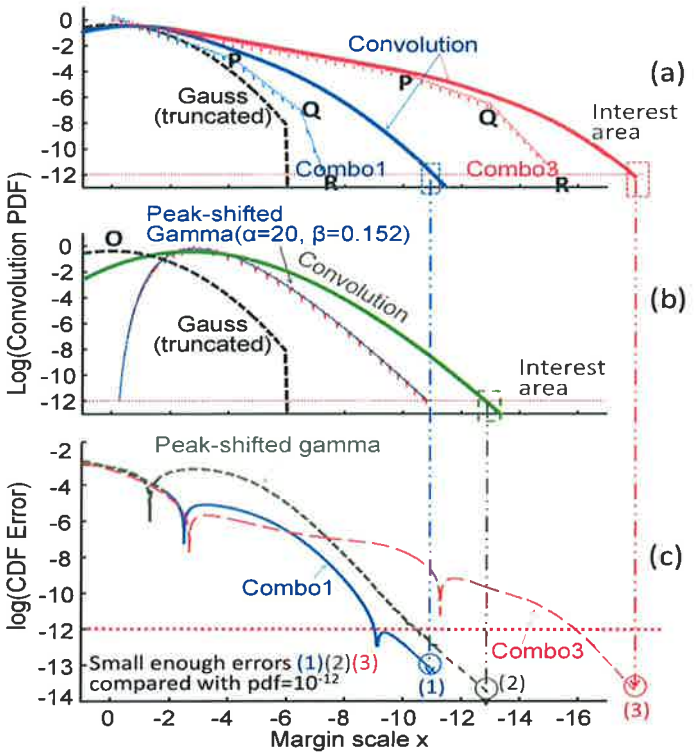


Figure 3.23: Convolutions of (a) 3-different sloped combined gamma and the truncated Gauss distribution, (b) peak-shifted gamma and (c) cdf-error comparisons.

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 Figures 3.23(a) and (b)

shows, while reducing the error of cdf to less than 1%, as shown in Figure 3.23(c). As can be seen in Figure 3.23, 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.

3.8 Conclusions

In this chapter, we show 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.

We have proposed 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. This chapter shows 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.

This chapter has 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.

3.9 References

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

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

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

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

[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, A. Oates, "The amplitude of random telegraph noise: Scaling implications", Reliability Physics Symposium (IRPS), IEEE International, pp.1.1-1.3, 2012.

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

[8]. K. Takeuchi, T. Nagumo, K. Takeda, S. Asayama, S. Yokogawa, K. Imai, Y. Hayashi, "Direct Observation of RTN-induced SRAM Failure by Accelerated Testing and Its Application to Product Reliability Assessment " , in Digest of IEEE Symposium on VLSI Technology 2010, pp. 189-190, 2010.

[9]. Moon, T.K, "The expectation-maximization algorithm" in Signal Processing Magazine, IEEE, Volume: 13, Issue: 6, pp. 47-60, 1996.

Chapter 4

Forward/Inverse Problem Analyses for RTN Long-Tail Distributions

In this chapter, how the statistical SRAM design analyses should be changed is discussed when: 1) the shift-amount of the time-dependent (TD) voltage margin variations (MV) after the screening test will become larger than that before and 2) the shapes of the MV distribution will change from the Gaussian to the complex mixtures of Gamma distributions. We discussed on the SRAM TD-MV analyses with not only the forward problem but also the inverse problem, i.e., deconvolution analyses. The proposed algorithm for the deconvolution to circumvent the issues caused by high-pass filtering behavior is discussed. Based on the proposed convolution/deconvolution design analyses, it has been shown for the first time that: 1) detecting the truncating point of the distributions of TD-MV by the screening test and 2) predicting the required the MV-shift-amount by the assisted circuit schemes to avoid the out of specs in the market during the life-time, etc, has become enabled based on the target specification.

4.1 Introduction

The guard band (GB) design for the static random access memory (SRAM) will become an unprecedentedly crucial challenge because the increased time-dependent (TD) margin variations (MV)-caused failures can't be predicated by the ordinary Gaussian-based convolution analyses any more. This stems from the facts that: 1) TD-MV, (i.e., unknown MV after shipped to the market), will become much larger than the non-TD-MV, (i.e., given MV based on the measurements), resulting in the TD-MV dominating over the whole MV. This leads to increased pressure to figure out the unknown factors by solving the inverse problem, although the SRAM designers are unfamiliar with such kind of methodology until now and 2) the tail distribution of the convolution results of TD-MV and non-TD-MV obeys no longer Gaussian but more complex mixtures of Gamma distributions [1]-[4], as shown in Figures 4.1(a) and 4.1(b).

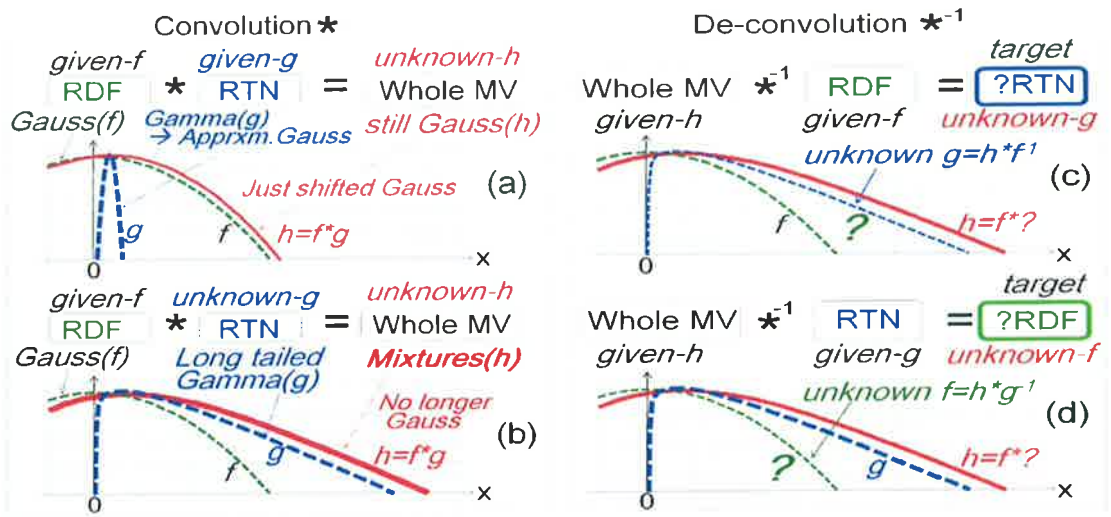


Figure 4.1: (a) Traditional, (b) RDF/RTN and its convolution, (c) And (d) deconvolution detecting the unknown factors.

The traditional SRAM statistical analyses including its convolution integral rely on the Gaussian model given its parameters extracted by the measured data. However, if the non-Gaussian unknown factors account for no longer just a fraction but a large percentage of the whole MV(shown in Figure 4.1(b)), this study has solved the non-Gaussian inverse problem based on the pre-defined hypothesis of the unknown factors or final target specifications, as shown in Figures 4.1(c)-4.1(d) and Figures 4.2(b)-4.2(d). To make clear the challenges behind the proposed ideas in this chapter, the concepts of what will be crucial in the coming process generations are shown more in detail in Figure. 4.2.

Figure 4.2(a) shows the concept for the required convolution methods to extrapolate the whole MV distribution based on the given data of the RDF and the RTN. The examples for the inverse problems are shown in Figures 4.2(b)-(d). Figure 4.2(b) recounts the following scenarios: target specs are predefined. The RTN distribution is also predefined or given, then, want to know the required truncating point (TP) of the minimum operating voltage (VDDmin) for the screening test. TP should be decided based on the RDF distributions to avoid any out of specs after shipped to the market. This can be solved by the de-convolution analysis. Figure 4.2(c) assumes the different scenario: both of the target specs and the RTN distribution are predefined or given, then want to know how much need to shift the whole MV by using the MV assisted circuit schemes (ASSTS)[5]-[6]. Another scenario is shown in Figure 4.2(d): the target specs and the truncated RDF distribution are predefined or given, then, want to specify the device target of reducing the amplitude of RTN.

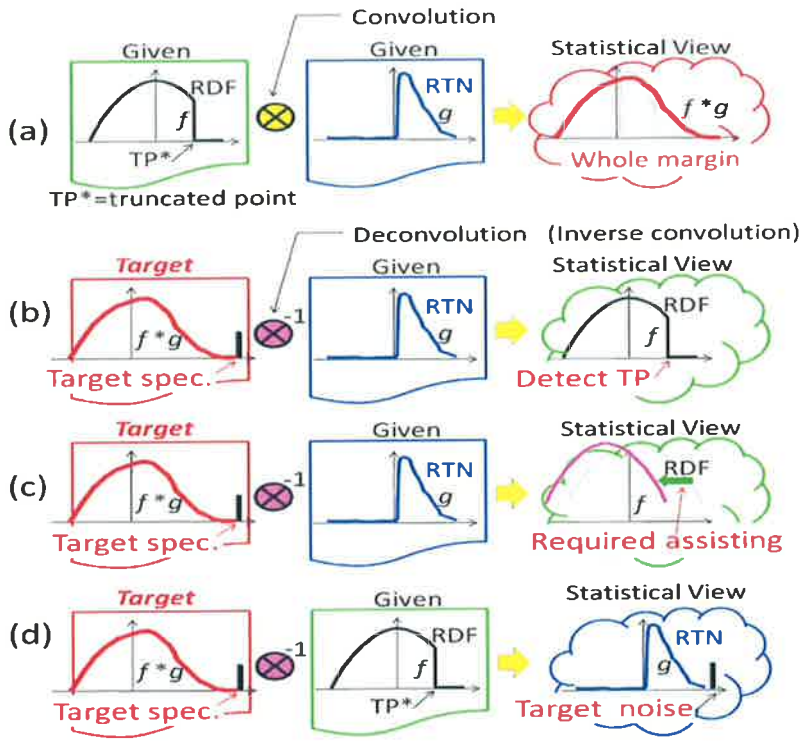


Figure 4.2: (a) Convolution of RDF and RTN, (b)-(d) examples of inverse problem required to detect the unknown factors.

Table 4.1: Comparisons between the conventional and proposed convolution/deconvolution.

(a) Convolution $h = f*g$

$h_1 <= f_1*g_1$
$h_2 <= f_1*g_2+f_2*g_1$
$h_3 <= f_1*g_3+f_2*g_2+f_3*g_1$
⋮ ⋮
$h_n <= f_1*g_n+f_2*g_{n-1}+f_3*g_{n-2}+f_4*g_{n-3}$ ····· $+f_n*g_1$

(b) De-convolution $g = h*f^{-1}$

$g_1 <= h_1/f_1$
$g_2 <= (h_2 - f_2*g_1)/f_1$
$g_3 <= (h_3 - f_2*g_2 - f_3*g_1)/f_1$
⋮ ⋮
$g_n <= (h_n - f_2*g_{n-1} - f_3*g_{n-2} - f_4*g_{n-3}$ ····· $- f_n*g_1)/f_1$

(c) Proposed De-convolution

Optimization problem $g = h*f^{-1}$	Objective funct. $O(g_n)$
g_1 :Min_{search}[$h_1 - O(g_1)$]	$(f_1:f_n) * g_1$
g_2 :Min_{search}[$(h_1:h_2) - O(g_2)$]	$(f_1:f_n) * (g_1:g_2)$
g_3 :Min_{search}[$(h_1:h_3) - O(g_3)$]	$(f_1:f_n) * (g_1:g_3)$
⋮ ⋮	⋮
g_n :Min_{search}[$(h_1:h_n) - O(g_n)$]	$(f_1:f_n) * (g_1:g_n)$

In order to make the proposed methods more easy to use at the factory field site, we use only the Excel tool for both of the convolution and deconvolution including the optimization process in the deconvolution. The expressions used for these calculations are shown in Table 4.1. The distributions are x-directly divided into the segmentation (# of segmentation is n). The RDF, RTN, and the convolution results are referred to f, g, and h, respectively.

4.1.1 Convolution Computing

The convolution can be given by the equation shown in Table 4.1(a). Since the convolution of f and g can be considered as the forward problem, if the mathematical equation and the parameters at the segmentation of f1-fn and g1-gn, are given, the observation of h1-hn can be easily computed. This operation plays a role of low-pass filtering, resulting in h=f*g a smoothed curve.

4.1.2 Issues Facing Deconvolution Operations

The deconvolution can be basically computed based on the equation shown in TABLE-1(b). If the observation of h1-hn and the parameter of f1-fn at the segmentation are given with the mathematical equation, the unknown parameter of g1-gn can be easily computed under the limited conditions.

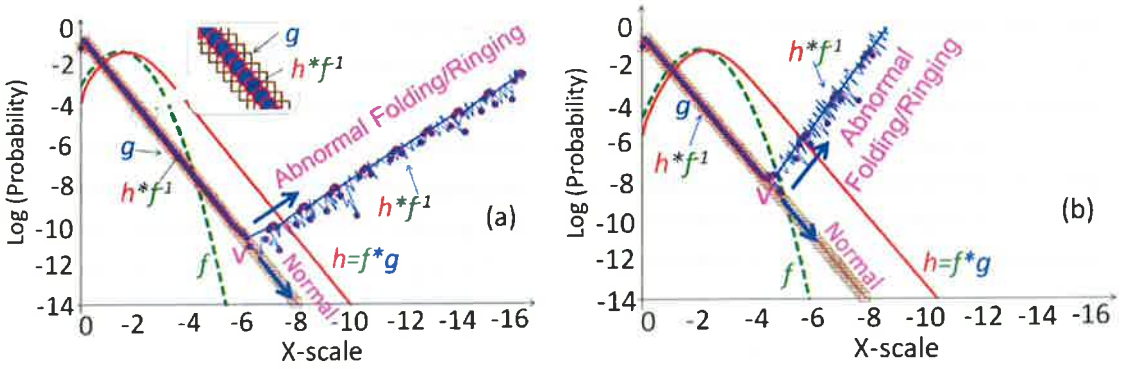


Figure 4.3: The abnormal phenomenon of V-shaped folding and ringing when using the conventional deconvolution algorithm. The difference between (a) and (b) is just the x-phase shift of f.

However, since the deconvolution has to be considered as the inverse problem, the invert operation, i.e., high-pass filtering, has to be executed, as shown in Table 4.1(b). This causes the issues facing the inverse operations: 1) division by zero and 2) Gibbs phenomenon, resulting in a ringing curve. As can be seen in Figure 4.3, the deconvolution of the RTN causes an abnormal folding and ringing, resulting in the RTN distribution

significant deviation from the expected curve. This stems from the required invert operation, i.e., high-pass filtering to do deconvolution. The point of V is very sensitive to the relationship of the gradients between the f, g and the h as can be seen when comparing those of Figures 4.3(a) and 4.3(b).

To address this issue, we try to circumvent the invert operation as much as possible and instead, solve it as the optimization problem to abstract the parameter RTN (g) of the objective function (partial integral of the h: sum(hi:h1, gi)), as shown in Table 4.1(c). The unknown parameter g1-gn can be searched so that the difference between the given observation data of h1-hn and the objective function O(gi) can be minimized.

4.2 Discussions on the Error of the Convolution/Deconvolution results

To verify the effectiveness of the proposed methodologies, the six types of RTN distributions are assumed. To normalize the ratio of the variation amplitude (ΔX) for RDF and RTN, the ΔX for RDF distribution is kept the same. It is assumed based on the references [1]-[4] that (1) the RTN distributions have several convex and concave folding points (O-P-Q-R-S) due to the variable atomistic variation factors, as shown in Figures 4.4(a)-4.4(f). Each slope (O-P, P-Q, Q-R, R-S) is represented by the shape parameter (β) of the Gamma distribution. In addition, to make clear the RDF-truncating effect on the convolution results, the tails of RDF (f) is truncated at the X=5. This truncating point (TP) gives impact on the critical VDDmin and the chip yield for the screening test before shipped to the market. The interesting zone for the fail-bit counts predictions (FBCP) is around the probability density function (pdf) of 10-12. This is because the pdf of 10-12 corresponds to the case of 99.9% chip yield of 1Gbit-chip.

4.2.1 Convolution Results

As can be seen in the Figures 4.4(a)-4.4(f), the convolution itself can be successfully done based on the equations of Table 4.1(a) even if the shape of the curves is a complex mixtures of the different sloped Gamma distributions. The results point out that the tails of the distributions of the convolution results (h=f*g) are no longer traditional Gaussian but strongly affected by the complex Gamma mixture tails (O-P-Q-R-S) of RTN (g).

4.2.2 New Deconvolution Results

The deconvolution results for the RTN(g) and RDF(f) are shown in Figures 4.5 and 4.6, respectively. The error of the deconvolution is defined as the relative error of the extrapolated value at each point, i.e., |gi-gdeconv|/gi, and |fi-fdeconv|/fi for RTN and RDF,

respectively. The g is the golden value for the comparison that is used when solving the forward problem of h=g*f. The gdeconv is solved by the inverse problem (deconvolution of given h and f). As can be seen in Figure 4.5, the relative errors of RTN deconvolution for (a), (b), and (c) are suppressed to small enough value of < 10-6, 10^{-8} and 10^{-8}, respectively. The errors of RDF deconvolution for all of (a),(b), and (c) in Figure 4.6 are suppressed to $<10^{-8}$. It is noted that the truncating point of RDF (disconnected point, i.e., jump to zero) is also well extrapolated by the proposed deconvolution algorithm while circumventing the ringing caused by the high-pass filtering.

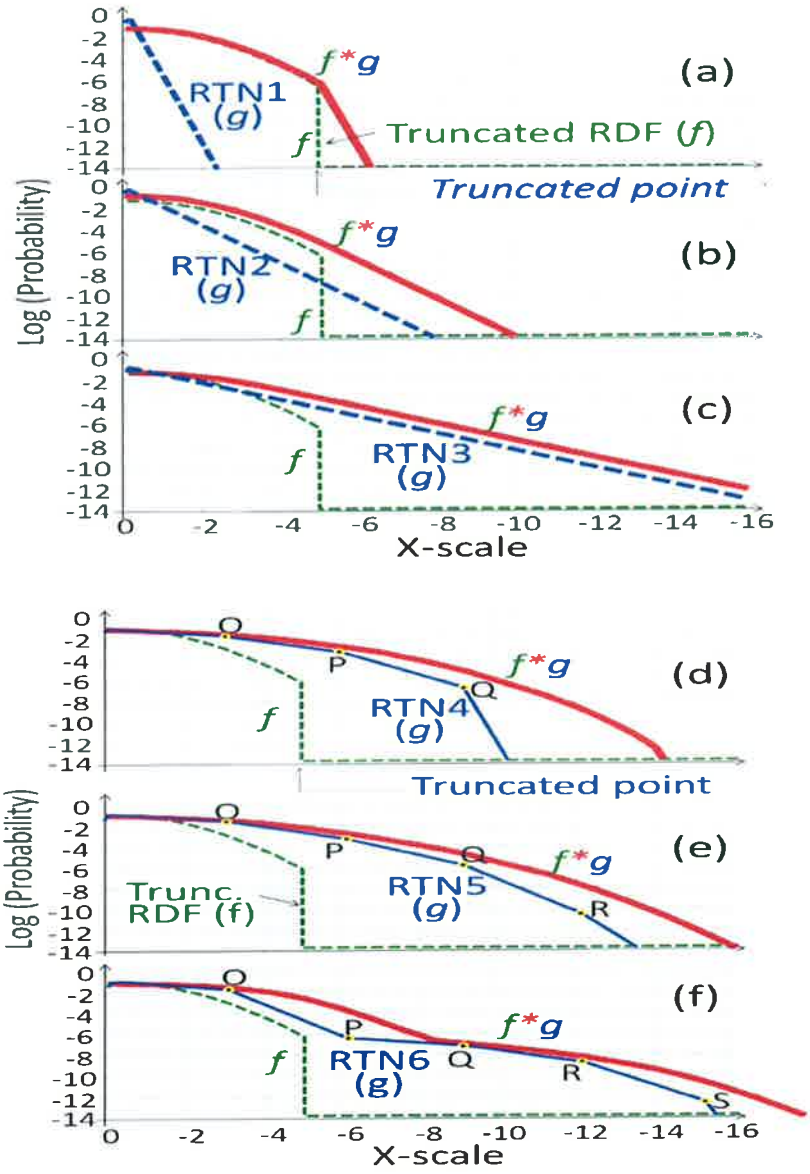


Figure 4.4: Slope of RTN(g) and the convolution(h=f*g), (a)-(c): different single slopes (d)-(f): different slope mixtures.

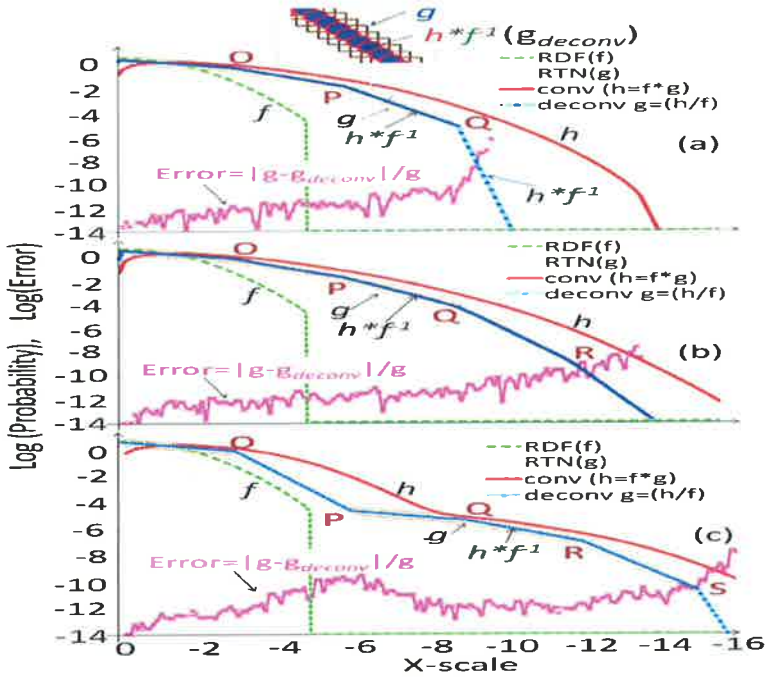


Figure 4.5: Tracing of the deconvolution gdeconv=h*f^{-1} comparing with the RTN(g). Relative error=|g-gdeconv|/g.

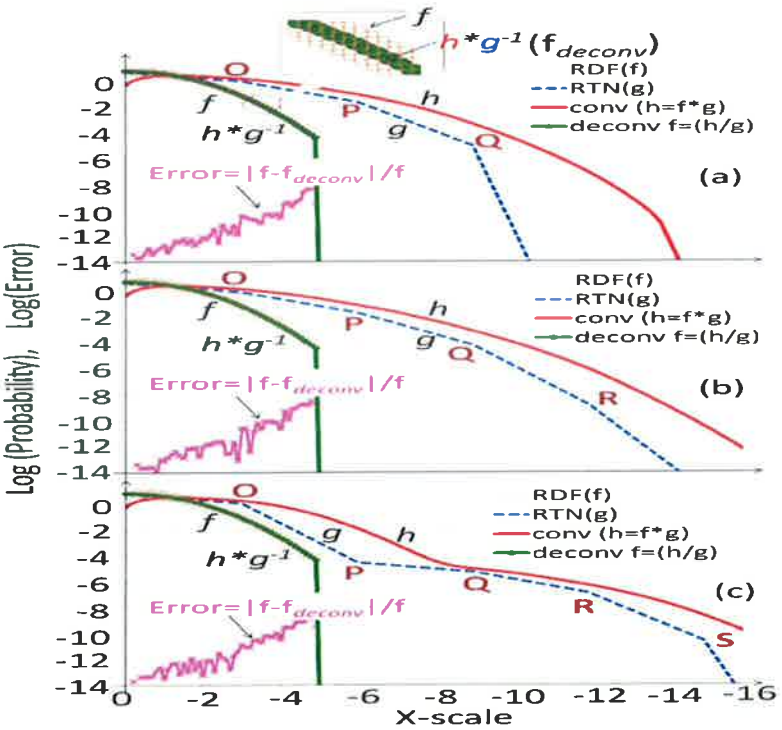


Figure 4.6: Tracing of the deconvolution fdeconv=h*g^{-1} comparing with the RDF(f). Relative error=|f-fdeconv|/f.

4.3 Detecting the Truncated Point

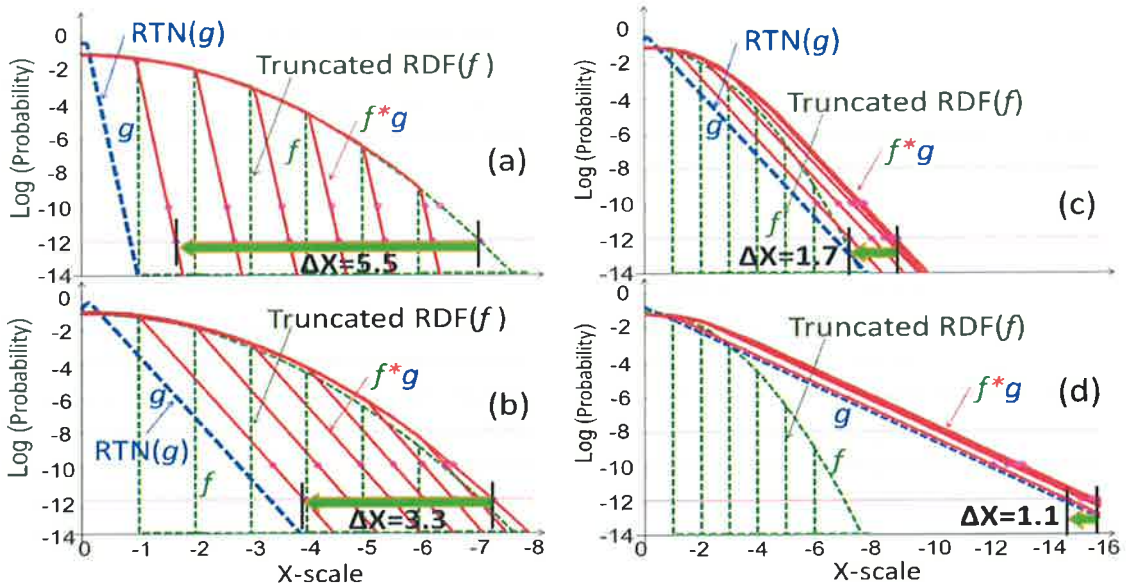


Figure 4.7: RDF-truncating point dependencies of the convolution results (f*g). (a)-(b): short tail RTN (smaller β), (c)-(d): long tail RTN (larger β).

As explained in Figure 4.2(b), the truncated point (TP) can be detected by the proposed methodology. The four types of the slope(β) of RTN(g) impacts on the convolution results (f*g) are shown in Figures 4.7(a)-4.7(d). It is found that the slope of the convolution tails (f*g) is strongly affected by the slope of the RTN(g) when the slope of the g is steeper than RDF(f), as shown in Figures 4.7(a) and 4.7(b). Contrary, the slope of g is shallower than that of the f, it gives much less impact on the convolution results (f*g), as shown in Figures 4.7(c) and 4.7(d).

The RTN-slope (β) dependencies of the shift amount (ΔX) of the convolution results (f*g) is shown in Figure 4.8. It should be noted that the shifting (ΔX) of the tails of the whole MV (f*g) will become harder as the RTN-slope becomes shallower, i.e., larger β It indicates that the ordinary truncating ways by the screening test can't work well any more for avoiding the out of specs after shipped to the market. This is because the operation of the convolution (f*g) does work as a low-pass filtering.

On the contrary, the operation of deconvolution works as a high-pass filtering; the sensitivity to detecting the truncated point (TP) is increased as the RTN-slope becomes shallower, i.e., larger β as shown in Figure 4.9. The sensitivity for β=0.56 becomes 5x larger than that for β=0.03. Since the β of RTN slope has upward trend as the device down-scaling, this operation becomes more important for the SRAM analyses.

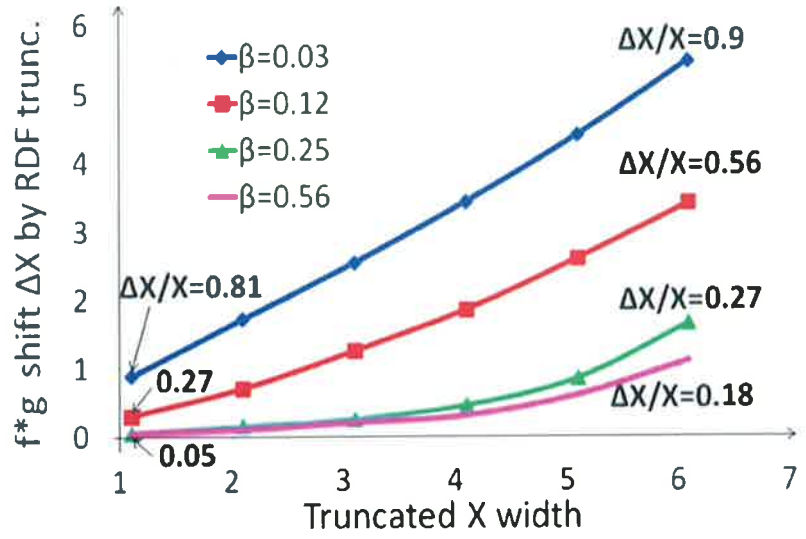


Figure 4.8: Slope (β) and truncated width X dependencies of the amount of f*g shift (ΔX).

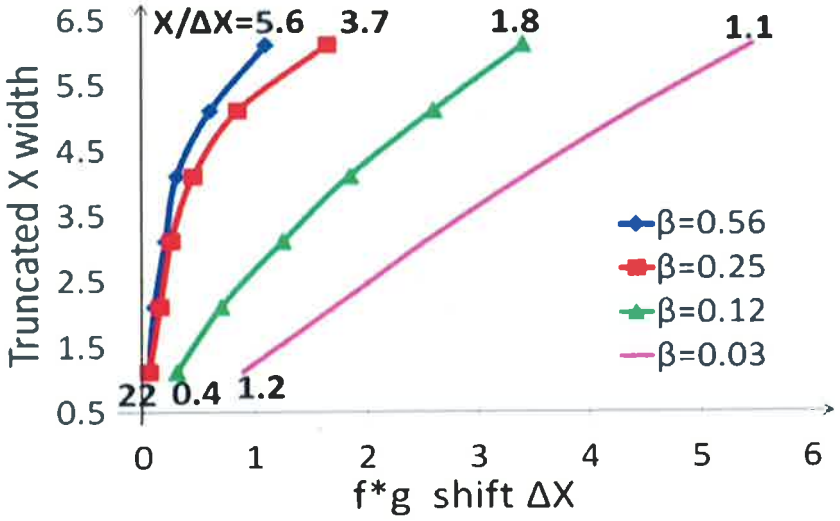


Figure 4.9: RTN-Slope (β) and f*g shift (ΔX) dependencies of the amount of the truncated width X extrapolated by the deconvolution.

4.4 Errors of Deconvolution

Figure 4.10 shows the RTN-slope dependencies of the relative errors of the RTN-deconvolution results. The amount of the error has a slope dependency but it is suppressed within 10^{-4} to 10^{-13}.

The RTN-slope dependencies of the relative errors of the RDF-deconvolution results are shown in Figure 4.11. The amount of the error doesn't have a slope dependency. The error is smaller than 10^{-7}.

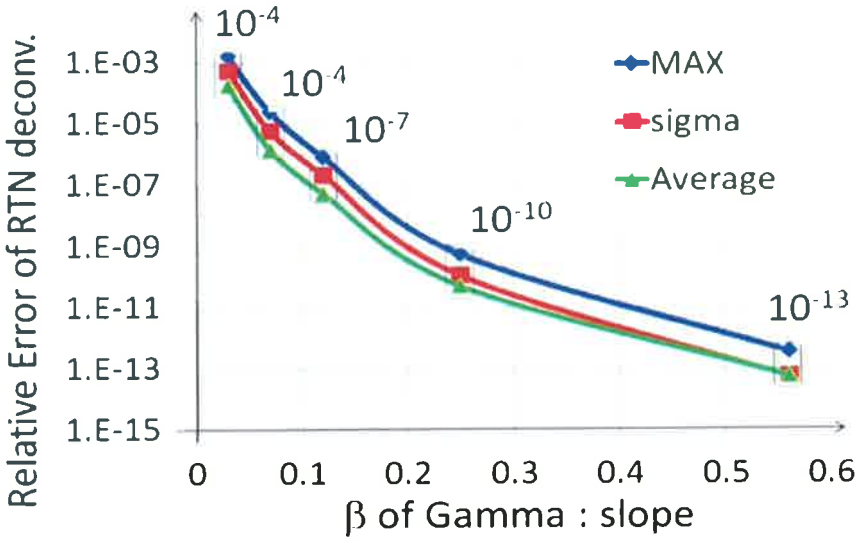


Figure 4.10: RTN-Slope (β) dependencies of the relative error of the RTN deconvolution.

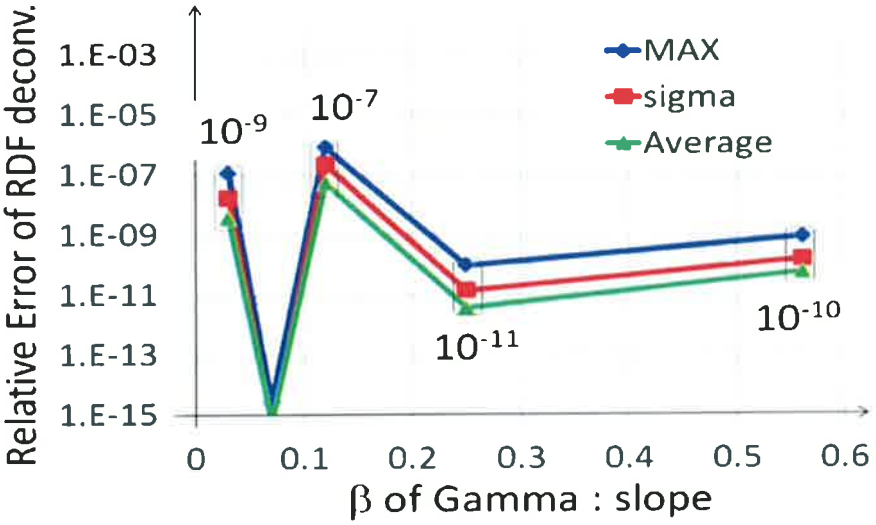


Figure 4.11: RTN-Slope (β) dependencies of the relative error of the RDF deconvolution.

4.5 Conclusions

In this chapter, we discussed, how the SRAM forward/inverse problems should be solved with the convolution/deconvolution analyses when the shift-amount of the TD-MV after the screening will become larger than that of before.

We demonstrated that the proposed methodologies can be used for solving the SRAM forward/inverse problems and can work well back-and–forth while suppressing the errors within the orders of 10^{-4} to 10^{-13} between the objective function (whole MV) and the given parameters (either of RTN and RDF).

We have shown that solving the optimization problem will be needed for getting 1) the unknown RTN, 2) the truncating points of RDF and 3) the required MV shift amount by the circuit assisted circuit schemes (ASSTS) to avoid any out of specs after shipped to the market. In order to circumvent the issue that the ordinary deconvolution can cause an abnormal folding and ringing errors.

We conclude that the ordinary MV truncating by the screening test to avoid the out of specs can't work well any more in the coming process generations. Contrary, the required truncating point to avoid the out of specs can be detected by solving the inverse problems between the target MV spec. and the given RTN distributions.

4.6 References

[1]. K. Takeuchi, T. Nagumo and T. Hase "Comprehensive SRAM Design Methodology for RTN Reliability", Symposium on VLSI Technology, pp. 130-131, June 2011.

[2]. K. Takeuchi, T. Nagumo, K. Takeda, S. Asayama, S. Yokogawa, K. Imai, Y. Hayashi, "Direct Observation of RTN-induced SRAM Failure by Accelerated Testing and Its Application to Product Reliability Assessment " , in Digest of IEEE Symposium on VLSI Technology 2010, pp. 189-190, 2010.

[3]. X. Wang, A.R. Broun, B. Cheng, A. Asenov, "RTS amplitude distribution in 20nm SOI FinFETs subject to Statistical Variability", Simulation of Semiconductor Processes and Devices SISPAD 2012, pp.296-299, 2012.

[4]. 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.

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

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

Chapter 5

Algebraic and Iterative Optimization Problem Based Deconvolutions

In this chapter, we make a comparative review of methods for the SRAM deconvolution analyses and discusses the pros and cons between the two methods of linear algebraic deconvolution (LAD) and the iterative optimization problem based deconvolution (IOPD) to extract the RDF and/or RTN from the complex mixtures of gauss and gamma distributions characterizing an overall nano-scaled SRAM margin variation. The study has demonstrated and shown for 1) the IOPD method can avoid the abnormal errors in both cases for the RDF and RTN deconvolution and 2) the LAD method can suppress the deconvolution error for the RDF more than IOPD but unfortunately confronts the issues on excessive deconvolution errors for the RTN.

5.1 Introduction

The designers for the SRAM will confront the unprecedentedly crucial issues on the GB design because the facts that: 1) time-dependent (TD) margin variations (MV) TD-MV, (i.e., unknown MV after the shipment), will govern over the non-TD-MV, (i.e., given MV based on the measurements). This causes to increase a pressure to figure out the unknown factors by solving the inverse problem, although the SRAM designers are still unfamiliar with it and 2) the tail distribution of the convolution results of TD-MV and non-TD-MV obeys no longer Gaussian.

The traditional SRAM statistical analyses rely on the Gaussian model and its parameters are extracted by the measured data. However, if the non-Gaussian unknown factors account for no longer just a fraction but a large percentage of the whole MV, we have to solve the non-Gaussian inverse problem based on the pre-defined hypothesis of the unknown factors or final target specifications, as shown in Figure 5.1. To make clear the challenges behind the proposed ideas in this chapter, the concepts of what will be crucial in the coming process generations are shown more in detail in Figure 5.1. Figure 5.1(a) shows the concept for the required convolution methods to extrapolate the whole MV

distribution based on the given data of the random dopant fluctuation (RDF) and the random telegraph noise (RTN). The examples for the inverse problems are shown in Figure 5.1(b).

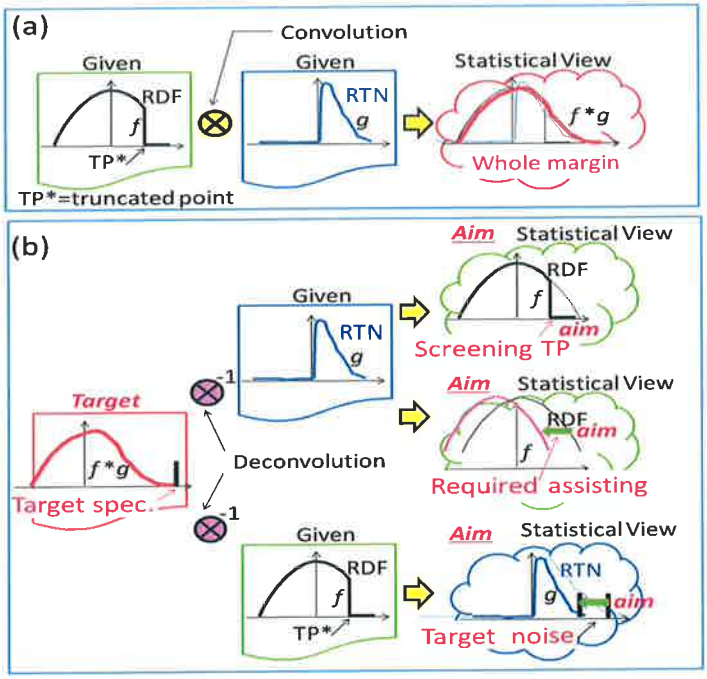


Figure 5.1: Concepts of (a) the convolution method and (b) deconvolution method and its applications

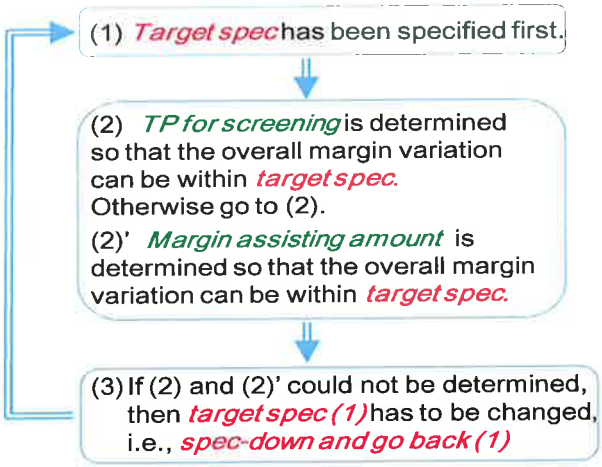


Figure 5.2: Relationships between the target spec and the TP for screening and margin assisting amount design.

Figure 5.1(b) recounts the following scenarios as shown in Figure 5.2: (1) target specs are predefined. (2) The RTN distribution is also predefined or given, then, want to know the required truncating point (TP) of the minimum operating voltage (VDDmin) for the

screening test. TP should be decided based on the RDF distributions to avoid any out of specs after shipped to the market. it can be used the deconvolution analysis for this purpose. It can be assumed the different scenario: both of the target specs and the RTN distribution are predefined or given, then want to know how much need to shift the whole MV by using the MV assisted circuit schemes (ASSTS)[5]-[6]. Another scenario is: the target specs and the truncated RDF distribution are predefined or given, then, want to specify the device target of reducing the amplitude of RTN. The target specs have to be changed unless improved overall margin variation can be put within the spec, as shown in Figure 5.2.

The traditional SRAM statistical analyses including its convolution integral rely on the Gaussian model given its parameters extracted by the measured data. However, if the non-Gaussian unknown factors account for no longer just a fraction but a large percentage of the whole MV, we should solve the non-Gaussian inverse problem based on the pre-defined hypothesis of the unknown factors or final target specifications.

5.2 Deconvolution of RDF and RTN

The convolution can be given by the equation shown in Table 5.1(a). Since the convolution of f and g can be considered as the forward problem, if the mathematical equation and the parameters at the segmentation of f1-fn and g1-gn are given, the observation of h1-hn can be easily computed. This operation plays a role of low-pass filtering, resulting in h=f*g a smoothed curve. On the contrary, the deconvolution is the process to extract the unknown one/two factor (f or/and g) from the overall distribution (h) given by the convolution of f with g. This usually requires the characteristics of the convolution (h) to be known. In this chapter, we compare two deconvolution methods. 1) Linear Algebraic Deconvolution (LAD), which provides a fast computation but the stability is often unsatisfactory. The deconvolution can be basically computed based on the equation shown in Table 5.1(b). If the observation of h1-hn and the parameter of f1-fn at the segmentation are given with the mathematical equation, the unknown parameter of g1-gn can be easily computed under the limited conditions. However, since the deconvolution has to execute the derivative operation, its behavior to the output is similar to the high-pass filtering. This causes the issues on abnormal errors facing the inverse operations: division by zero and Gibbs phenomenon. 2) Iterative optimization problem based deconvolution (IOPD), where the deconvolution can be computed based on the equation shown in Table 5.1(c). If the observation of h1-hn and the objective parameter O(gn) are given (O(gn) is the convolution of gn and f), the unknown parameter of g1-gn can be determined by searching the parameter set g1-gn, which can minimize the difference between seg(h1: hn) and O(gn).

Where, seg(h1: hn) denotes the line segmentation connecting the points from h1 to hn. Since IOPD method can be considered as the forward problem, this allows avoiding the issues facing the inverse operations of LAD method, as it was already mentioned.

Table 5.1: Comparisons between the linear algebraic deconvolution and iterative optimization problem based deconvolution

(a) Convolution h= f*g

$h_1 <= f_1*g_1$
$h_2 <= f_1*g_2+f_2*g_1$
$h_3 <= f_1*g_3+f_2*g_2+f_3*g_1$
⋮ ⋮
$h_n <= f_1*g_n+f_2*g_{n-1}+f_3*g_{n-2}+f_4*g_{n-3}$ ····· $+f_n*g_1$

(b) De-convolution g= h*f^{-1}

$g_1 <= h_1/f_1$
$g_2 <= (h_2 - f_2*g_1)/f_1$
$g_3 <= (h_3 - f_2*g_2 - f_3*g_1)/f_1$
⋮ ⋮
$g_n <= (h_n - f_2*g_{n-1} - f_3*g_{n-2} - f_4*g_{n-3}$ ····· $- f_n*g_1)/f_1$

(c) Proposed De-convolution

Optimization problem g= h*f^{-1}	Objective funct. $\boldsymbol{O}(g_n)$
g_1 :Min_{search}[$h_1 - \boldsymbol{O}(g_1)$]	$(f_1:f_n) * g_1$
g_2 :Min_{search}[$(h_1:h_2) - \boldsymbol{O}(g_2)$]	$(f_1:f_n) * (g_1:g_2)$
g_3 :Min_{search}[$(h_1:h_3) - \boldsymbol{O}(g_3)$]	$(f_1:f_n) * (g_1:g_3)$
⋮ ⋮	⋮
g_n :Min_{search}[$(h_1:h_n) - \boldsymbol{O}(g_n)$]	$(f_1:f_n) * (g_1:g_n)$

The concepts of what will happen at that time are shown in Figure 5.3. Figure 5.3 illustrates the probability density functions for RDF, RTN1 (40nm), RTN2 (<16nm) and RTN3 (<7nm), and its convolution results, respectively. In this study, we use 3 case of RTN to comparison between LAD and IOPD methods.

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. 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 Figure 5.3.

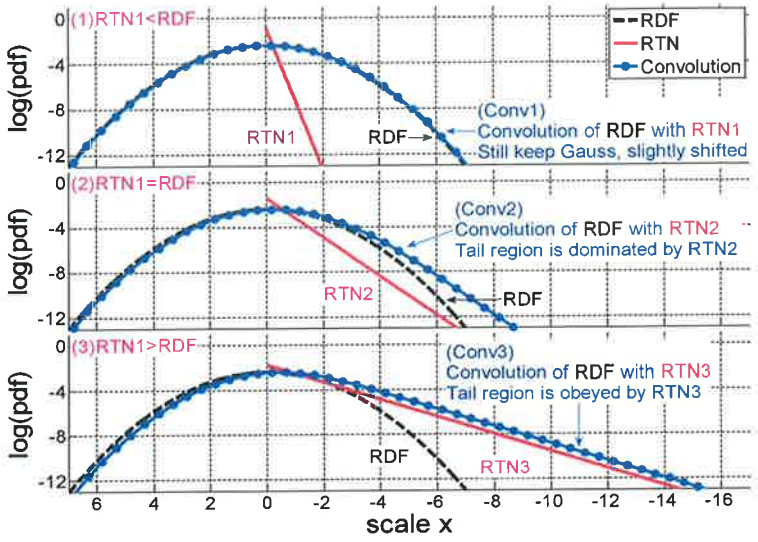


Figure 5.3: Comparisons of convolution results of (1) RTN<RDF, (2) RTN=RDF and (3) RTN>RDF.

5.3 RDF Deconvolution

One of the challenges in this study is to address how to know the required TP of VDDmin. As mentioned above, TP is based on the RDF distributions to avoid any out of specs after shipped to the market. In this section, have to point out that it is possible to extract RDF from the complex mixtures distributions.

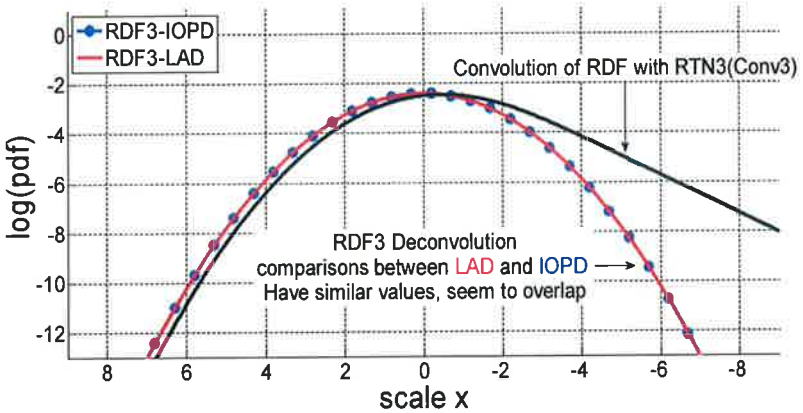


Figure 5.4: Examples, RDF3 deconvolution from Conv3, Comparisons between LAD and IOPD methods, the result seem to overlap.

Figure 5.4 shows the example of the comparisons of RDF3 deconvolution between LAD and IOPD methods. Where, RDF3 is the extracted distribution by the deconvolution of the overall distribution (Conv3 shown in Figure 5.3) with the RTN3 distribution. Both in LAD and IOPD methods, the deconvolution itself can be successfully done based on the

equations of Table 5.1(b)-(c) without any troubles. However, to verify the effectiveness of the two proposed methodologies, the relative errors are compared in the overall range of x-scale (i.e., the number of the sigma for Gaussian).

The extrapolated values at each point are shown as the relative error, i.e., |gi-gdeconv|/gi, and |fi-fdeconv|/fi for RTN and RDF, respectively. The g is the golden value for the comparison that is given by solving the forward problem of h=g*f. The gdeconv is the extracted data based on the deconvolution of h with f.

The relative error comparisons for the 3-cases of RDF1, RDF2, and RDF3 are shown in Figure 5.5. The LAD method suppresses the relative error for RDF1, RDF2, and RDF3 to small enough value of $< 10^{-14}$, 10^{-10} and 10^{-6}, respectively. It is worth noting that (1) LAD cans suppress the error more than IOPD. However it has a strong dependency of the slope of the RTN, i.e., the error for RDF3 is larger than that for RDF1, and (2) LAD has another dependency, i.e., as the x-scale is larger, the errors are increased. It should be noted that its dependencies becomes stronger as the slope of RTN becomes longer, i.e., RTN1<RTN3.

Generally, the overall of IOPD relative error is a little bit higher than those of LAD but the error distribution curve is more smoothed than LAD. Compared with LAD, IOPD doesn't have too many dependencies on the x-scale, while it still has the RTN slope dependency.

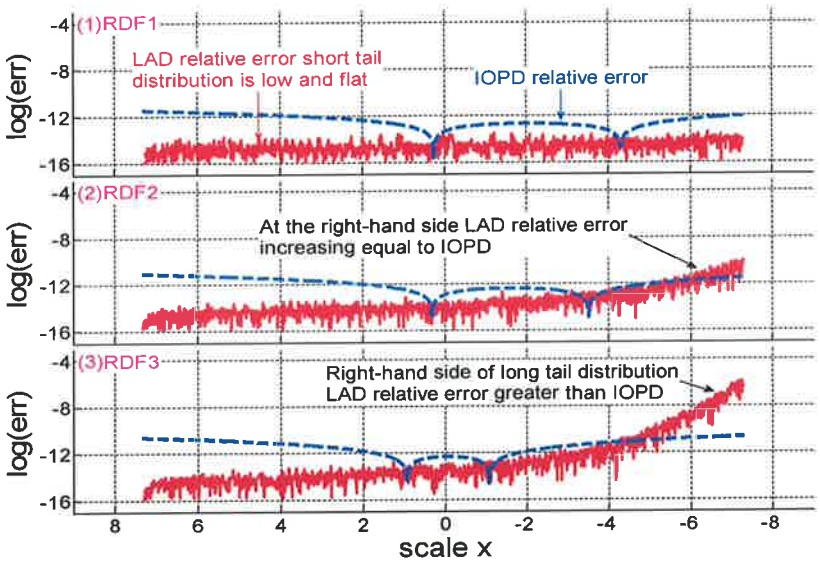


Figure 5.5: Comparisons of the relative error between LAD and IOPD methods of (1)RDF1, (2)RDF2 and (3)RDF3.

When discussing the screening conditions for volume production, some expected yield should be assumed. We assumed the target here that fail-probability is 10^{-12} to realize 99.9% yield of 1Gbit memory chip, which is quite realistic target. For instance, the points

of the x-scale of around x=6.5, x=8 and x=14 in Fig. 3 correspond to the fail-probability of 10-12 for Conv1, Conv2, and Conv3, respectively.

Figure 5.6 shows the comparisons of the cumulative density function (cdf) errors between LAD and IOPD. The 3-cases of the cdf errors of the convolution results of RDF with RTN are shown, respectively. The cdf errors for the "LAD" are smaller than that for the "IOPD" in the interest area, where fail-bit count analyses should account for the rare event. Overall, it can be said that the LAD provides a smaller errors of the RDF deconvolution. It can be applied to define truncating point (TP) for the screening test, and it can be used for guarantee of the yield-loss after shipped to the market due to the time-dependent RTN-caused failures.

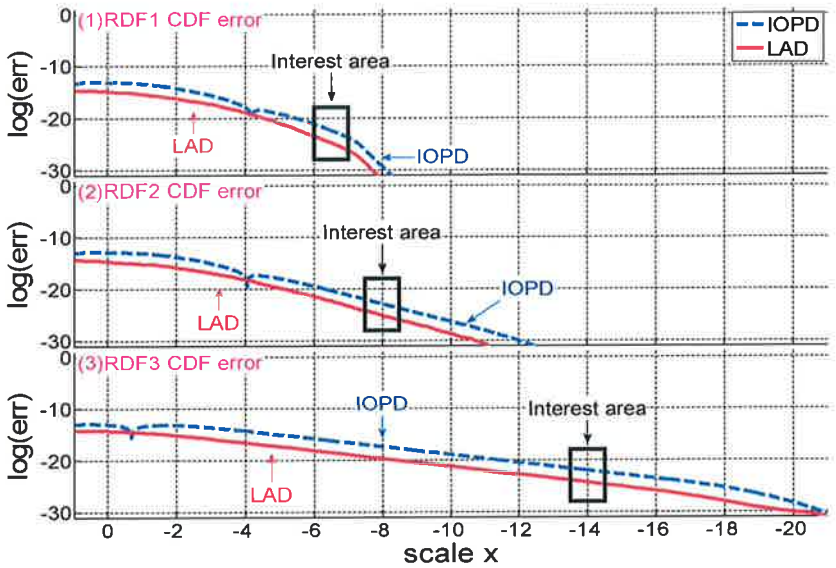


Figure 5.6: Comparisons of the cdf error between LAD and IOPD methods of (1) RDF1, (2) RDF2 and (3) RDF3.

5.4 RTN Deconvolution

In previous section, the RTN was predefined and we already know the RDF. In another way, if the target specs and the truncated RDF distribution are predefined, then, it can be specified the device target of RTN, as explained in Figure 5.2.

Figure 5.8 shows the error comparisons of the RTN deconvolution between LAD and IOPD methods. In this case, the target specs and RDF are predefined; IOPD method can extract the RTN without any troubles. On the contrary, the deconvolution of the RTN based on LAD method causes an abnormal folding and ringing, resulting in the RTN distribution significant deviation from the expected curve. The point of V is very sensitive to the relationship of the gradients between the f, g and the h as can be seen in Figure 5.7.

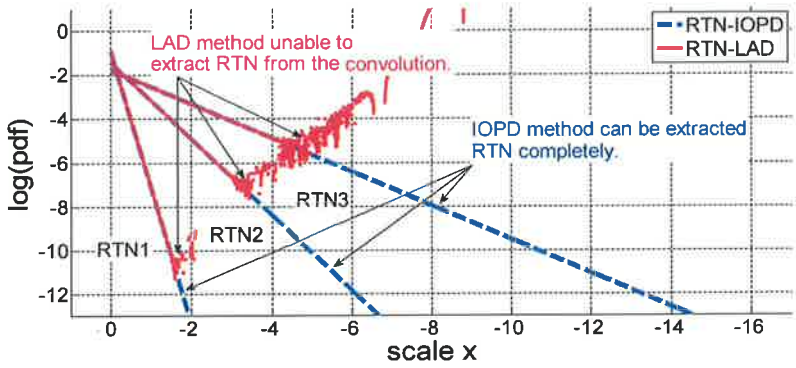


Figure 5.7: Comparisons of RTN deconvolution between LAD and IOPD methods of RTN1, RTN2 and RTN3.

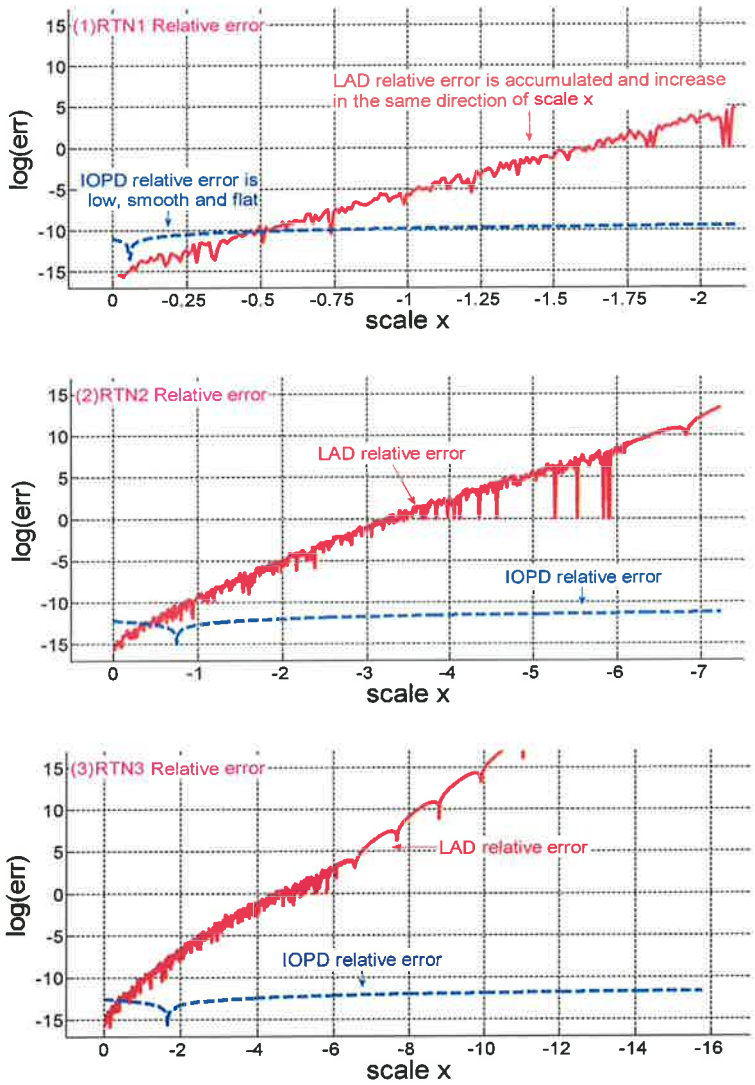


Figure 5.8: Comparisons of the relative error between LAD and IOPD methods of (1)RTN1, (2)RTN2 and (3) RTN3.

Based on these results shown in Figure 5.7, it can be said that the LAD method cannot be applied to extract the RTN from the complex mixtures. The relative error of RTN deconvolution for 3 cases of (1) RTN1, (2) RTN2 and (3) RTN3, are compared between LAD and IOPD methods. The errors for LAD method are excited to increase at a certain point and started to exponentially accumulate the error as the x-scale is increased, resulting in making V-shaped error distribution. This is caused by the problematic operation (derivative operation including division by zero) of inverse problem. It makes it more sensitive to error/noise, resulting in unstable operation. It becomes more difficult to use for SRAM RTN effect analysis with a priori information.

To illustrate the effects of the convolution/deconvolution of each method in a simple way, we have demonstrated that the proposed methodologies can be work well back-and–forth. In order to characterize the error of each convolution result, the "golden", which is given by the numerical calculation of convolution of RDF and RTN distributions, is used as a reference, as shown in Figure 5.3.

The results for the convolution/deconvolution when using LAD and IOPD methods are shown in Figure 5.9 and 10, respectively. It is found that the LAD method confronts the issue of V-shaped abnormal error and its folding point of V-shaped distribution depends on the RTN tail length. The deconvolution error for RTN3 is larger than that for RTN1, as shown in Figure 5.9. On the other hand, the IOPD method solves successfully the issues problematic deconvolution for all the cases of RTN tail length, as shown in Figure 5.10. This method can be an alternative method for stable deconvolution means.

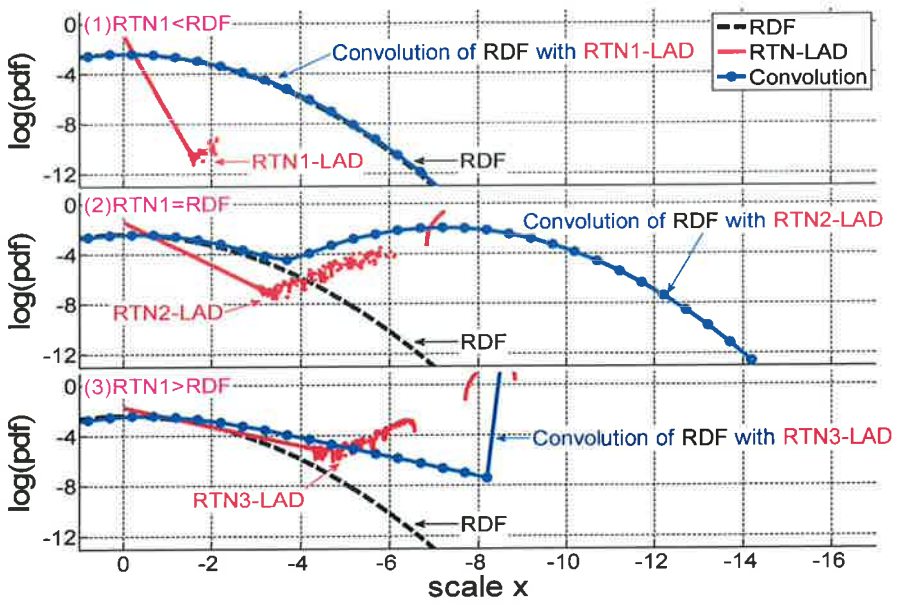


Figure 5.9: Comparisons of the convolution result between RDF with (1)RTN1-LAD, (2)RTN2-LAD and (3)RTN3-LAD.

The numbers of fail-bit count (FBC) errors for each method are compared, as shown in Figure 5.11. The FBC error is defined as the difference in the cdf value at a certain point between the "golden" and that for each means.

The comparison of "which is better" depends on the interest area, as shown in Figure 5.11. It can be said that the IOPD provides a smaller errors of the RTN deconvolution. It can be applied to define the target spec to avoid the out-of-spec in the market due to the time-dependent RTN-caused failures after the shipment.

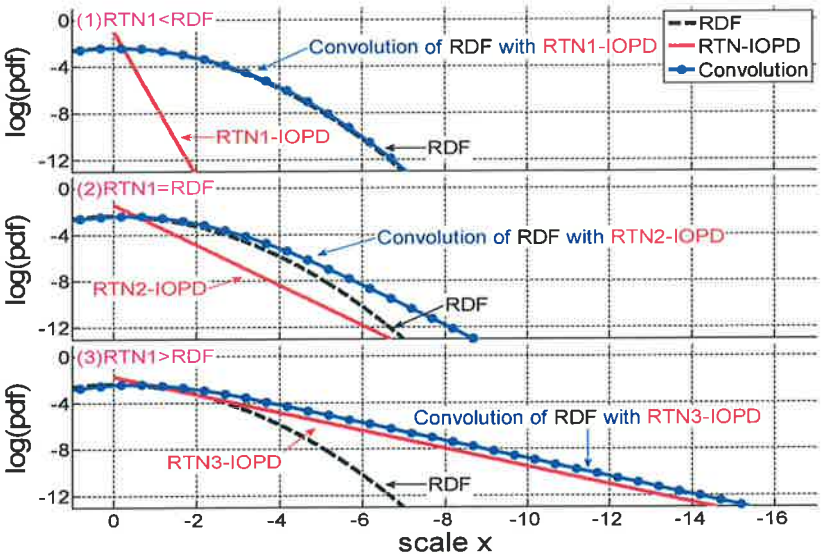


Figure 5.10: Comparisons of the convolution result between RDF with (1)RTN1-IOPD, (2)RTN2-IOPD and (3)RTN3-IOPD.

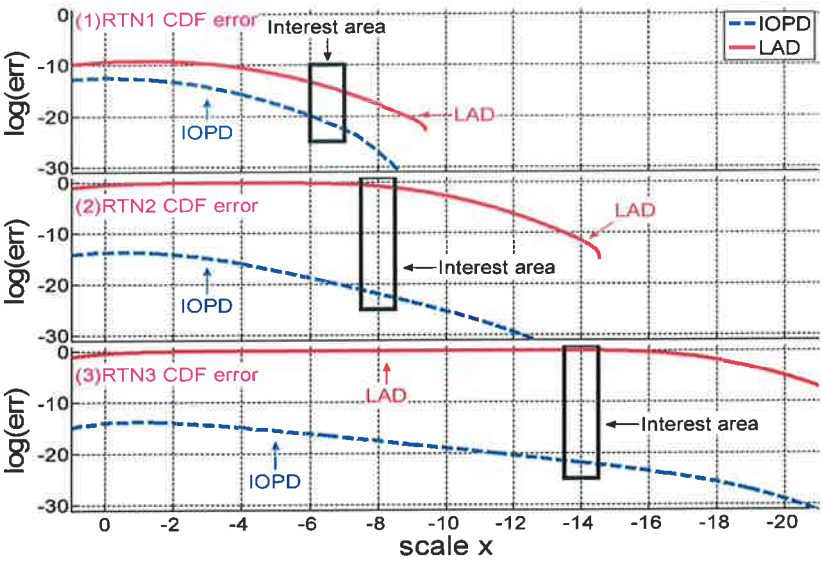


Figure 5.11: Comparisons of the cdf error between LAD and IOPD methods of (1)RTN1, (2)RTN2 and (3)RTN3.

5.5 Faster Convergence Algorithm for Solving Deconvolution

In this section, we propose a new deconvolution algorithm enabling a more stable and faster convergence in the optimization process for solving mixtures of gamma distributions. The algorithm reduces the number of parameters to be sought at once in the optimization problem, resulting in not only a decrease in required iteration cycles for convergence but also an increase in the accuracy of the deconvolution. To the best of our knowledge, it is for the first time to demonstrate the issues facing the iterative optimization problem based deconvolution algorithm (IOPD) shown in Figure 5.12(a). The proposed algorithms of iterative optimization problem based deconvolution with least mean square (IOPLSD) [7] are shown in Figure 5.12(b).

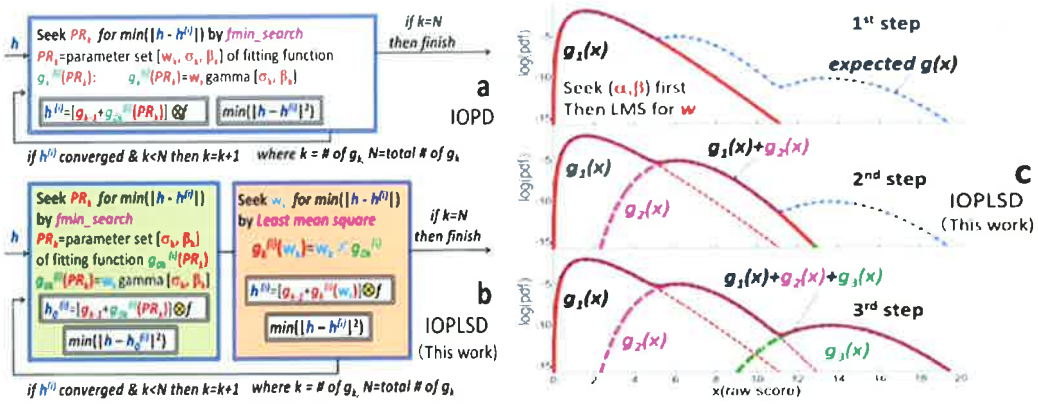


Figure 5.12: (a) IOPD algorithm, (b) IOPD with least means square (LMS) (IOPLSD) and (c) Step by step deconvolution for each peaked distribution.

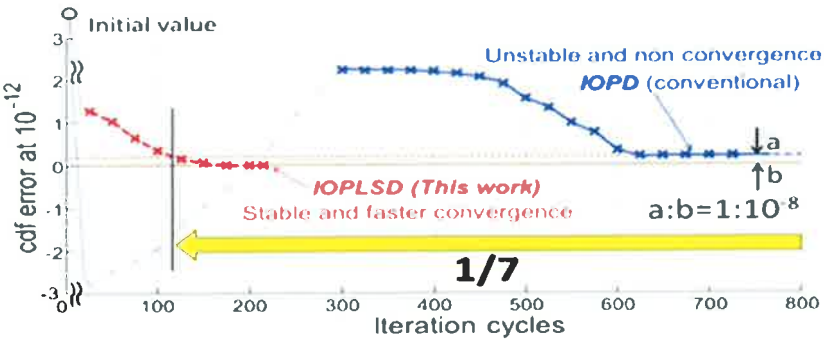


Figure 5.13: Comparisons of cdf error between IOPD and proposed IOPLSD.

The proposed IOPLSD introduces the two key ideas to reduce the number of fitting parameters at once from 9 to 2: (1) counting the peaks (#=3) and dividing into 3-stages and (2) 2-for-3 split of the 3-paremeters (wk, α, β) by adding the least mean square into two-groups of fmin-search for (α, β) and least-mean-square for (wk) (see Figure 5.12(b)). By

this algorithm, the number of parameters to be sought at once in optimization problem is reduced from 9 to 2. This allows a stable and faster convergence behavior as shown in Figure 5.12(c).

Figure 5.13 clearly shows that IOPLSD can provide much better characteristics of both (1) >7x convergence speed of iteration process for error reduction and (b) 10^8 x higher attained accuracy level of h'(x), compared with the conventional IOPD.

5.6 Conclusions

In this chapter, we made a comparative review of the two types of methods for convolution/deconvolution and discussed the pros and cons between the two methods of linear algebraic deconvolution (LAD) and the iterative optimization problem based deconvolution (IOPD). The comparisons have been done based on the case study when extracting the RDF and/or RTN unknown factors from the given complex mixtures of gauss and gamma distributions characterizing an overall nano-scaled SRAM margin variation.

We demonstrated that the IOPD method can avoid the abnormal errors well in both cases for RDF and RTN deconvolution, and the LAD method can suppress the deconvolution error for the RDF more than IOPD but unfortunately confronts the issues on the excessive deconvolution errors when extracting the RTN.

5.7 References

[1]. K. Takeuchi, T. Nagumo and T. Hase "Comprehensive SRAM Design Methodology for RTN Reliability", Symposium on VLSI Technology, pp. 130-131, June 2011.

[2]. K. Takeuchi, T. Nagumo, K. Takeda, S. Asayama, S. Yokogawa, K. Imai, Y. Hayashi, "Direct Observation of RTN-induced SRAM Failure by Accelerated Testing and Its Application to Product Reliability Assessment " , in Digest of IEEE Symposium on VLSI Technology 2010, pp. 189-190, 2010.

[3]. X. Wang, A.R. Broun, B. Cheng, A. Asenov, "RTS amplitude distribution in 20nm SOI FinFETs subject to Statistical Variability", Simulation of Semiconductor Processes and Devices SISPAD 2012, pp.296-299, 2012.

[4]. 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.

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

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

[7]. W. Somha , H. Yamauchi and M. Yuyu, "More Stable and Faster Convergence Algorithm for Solving Deconvolution Problem of Peak-Shifted Gamma Mixtures RTN Distributions ", The 29th International Technical Conference on Circuit/Systems Computers and Communications (ITC-CSCC, Phuket, Thailand) , pp.504-507, 1-4 July 2014.

Chapter 6

Iterative and Adaptively Segmented Forward Problem Based Non-Blind Deconvolution

This chapter proposes a ringing-error-free non-blind deconvolution technique featuring an iterative and adaptively segmented forward-problem based deconvolution (IASDCN) process. Unlike the algebraic based inverse operations, this eliminates any operations of differential and division by zero to successfully circumvent the issue on an abnormal V-shaped error. This effectiveness has been demonstrated for the first time with applying to a real analysis for the effects of the RTN and/or RDF on the overall SRAM margin variations. It has been shown that the proposed IASDCN technique can reduce its relative errors of RTN deconvolution by 10^{13} to 10^{15} times, which are good enough for avoiding the abnormal ringing errors in the RTN deconvolution process. This enables to suppress the cdf error of the convolution of RTN with RDF (i.e., fail-bit-count error) to $1/10^{10}$ error for the conventional algorithm.

6.1 Introduction

The GB design for the SRAM will become an unprecedentedly crucial challenge because the increased time-dependent (TD) margin variations (MV)-caused failures can't be predicated any more by only the convolution analyses [1-5]. This stems from the facts that TD-MV, (i.e., unknown MV after shipped to the market), will become much larger than the non-TD-MV, (i.e., given MV based on the measurements), resulting in the TD-MV dominating over the overall MV. This leads to an increased pressure to figure out the unknown factors by solving the inverse problem [6-8], although the SRAM designers are unfamiliar with such kind of methodology until now.

To make clear the background behind the proposed ideas in this chapter, the concepts of what will be crucial in the coming process generations are shown first in detail in Figure 6.1. Figure 6.1 shows the concept for the required convolution/deconvolution methods. The example for the inverse problem is shown in Figure 6.1(b) compared with the forward problems (shown in Figure 6.1(a)). Figure 6.1(b) recounts the following scenarios: the

target specs are predefined. The RTN distribution is also predefined or given in advance, then, have to know the required truncating point (TP) of the minimum operating voltage (VDDmin) for the screening test. The TP should be decided based on the RDF distributions to avoid any out of specs after the shipment to the market. This has to be solved by the de-convolution analysis. Unlike the convolution, the deconvolution is sort of ill-posed problem and troublesome operation [7-8]. For example, have to address the issues on an abnormal ringing error confronting the algebraic based inverse problem process [7-8].

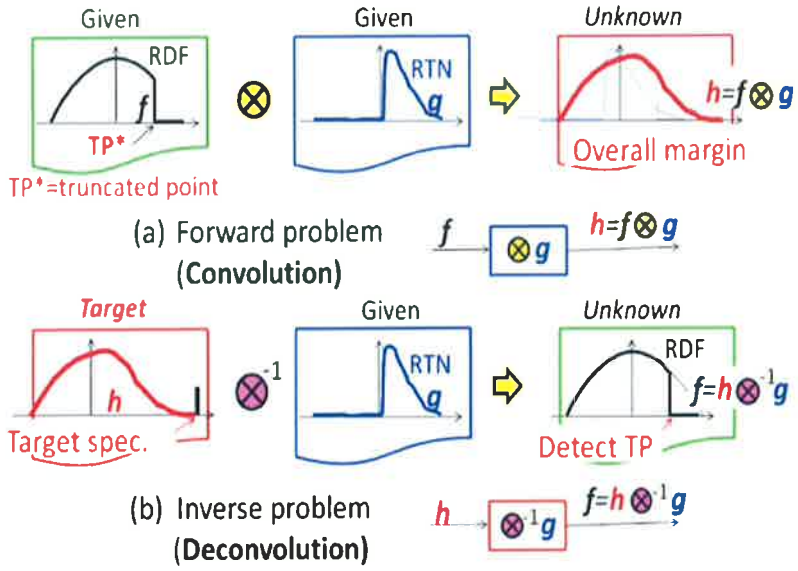


Figure 6.1: (a) forward problem and (b) inverse problem.

This abnormal error is caused by: (1) the division by zero and Gibbs phenomenon, resulting in a ringing and folding V-shaped curve. As a result, the deconvoluted RTN distribution is significantly deviated from the expected curve, as shown in Figure 6.2(a). This stems from the required invert operation, i.e., differentiation and feedbacks, resulting in accumulating of the errors, as shown in Figure 6.2(b). Once the accumulated error level exceeds a certain level, the ringing are excited and the error is amplified, as shown in Figure 6.2(a). This sort of high-pass filtering behavior makes the output more sensitive to the noise and the error, resulting in an unstable operation [7-8].

The purpose of this work is to enable: 1) to substantially avoid the abnormal V-shaped ringing errors by eliminating the need of the inverse operation, 2) to increase the deconvolution accuracy in the tail region. Since the SRAM fail probability is extremely rare-event level (pdf<10^{12}), the degree of precision for the tail distribution effects on the accuracy of the fail-bit count (FBC) prediction. Unlike the conventional optimization problem, which tends to ignore the rare-event probability zone, the proposed idea tries to

keep a sharp eye on the rare event probability area by introducing the segmented optimization and 3) to guarantee the good enough deconvolution precision even if the RTN distribution comprises the complex gamma mixtures with the multiple convex and concave folding points.

To the best of our knowledge, this is the first time to present the deconvolution algorithm for SRAM-designs featuring an iterative and adaptively segmented forward-problem based deconvolution (IASDCN) process and enabling to achieve the above mentioned three objects.

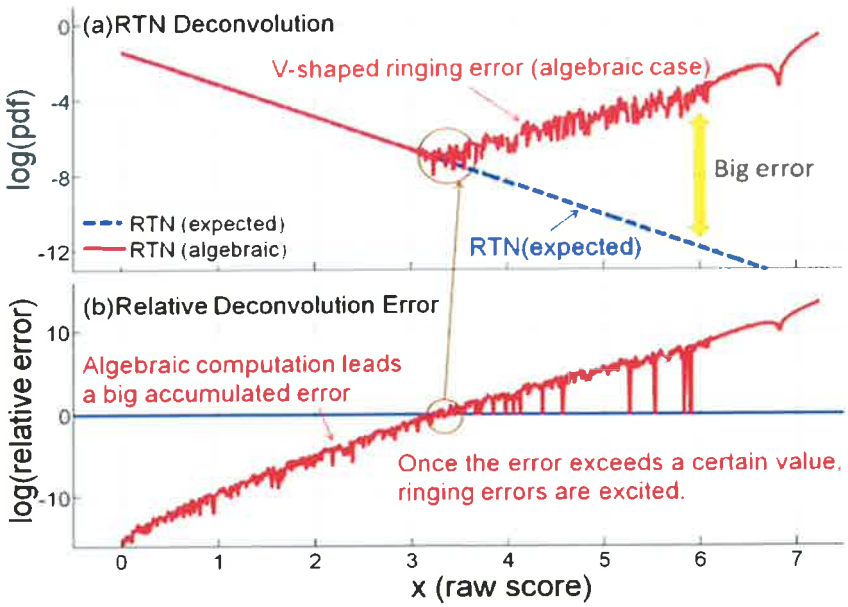


Figure 6.2: Issue of V-shaped error in deconvoluion process using algebraic way where division by zero can happen.

6.2 Proposed Iterative Deconvolution Algorithm

The key feature of the proposed deconvolution algorithm is not to use the inverse operation such as differentiation and division by zero at all (shown in Figure 6.3(b)). Instead, solving the optimization problem that seeks $f^{(i)}$ for minimizing ($|h\text{-}h^{(i)}|$), where h(i) is the convolution of $f^{(i)}$ with g, as shown in Figure 6.3(c). By avoiding the inverse operation, the behavior of the deconvolution process becomes insensitive to the noise and the error like the low-pass filtering. As a result, the accumulated deconvolution errors are significantly suppressed and its error level is reduced by $>10^{17}$ at raw score x=6, as shown in Figure 6.4(b). Any ringing noises and V-shaped errors are not observed any more when using the proposed deconvolution algorithm.

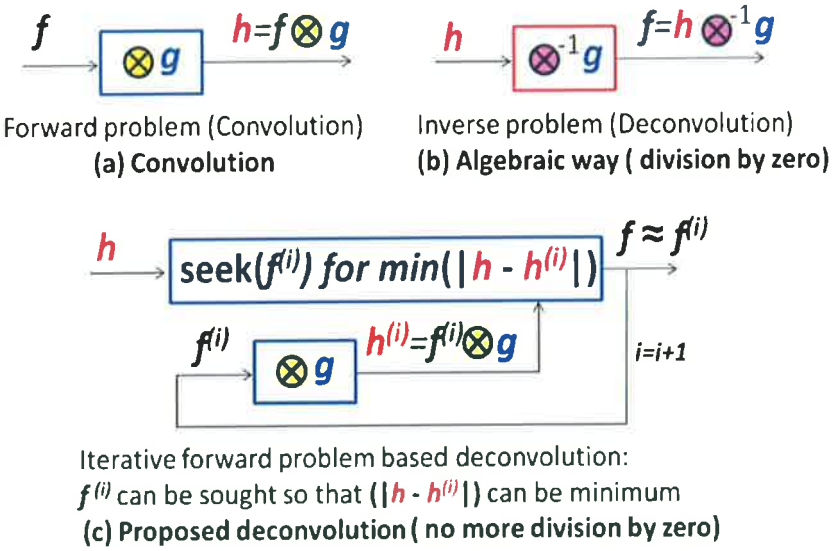


Figure 6.3: Deconvolution comparisons between (b) the conventional and (c) proposed one that uses only (a) convolution instead of (b).

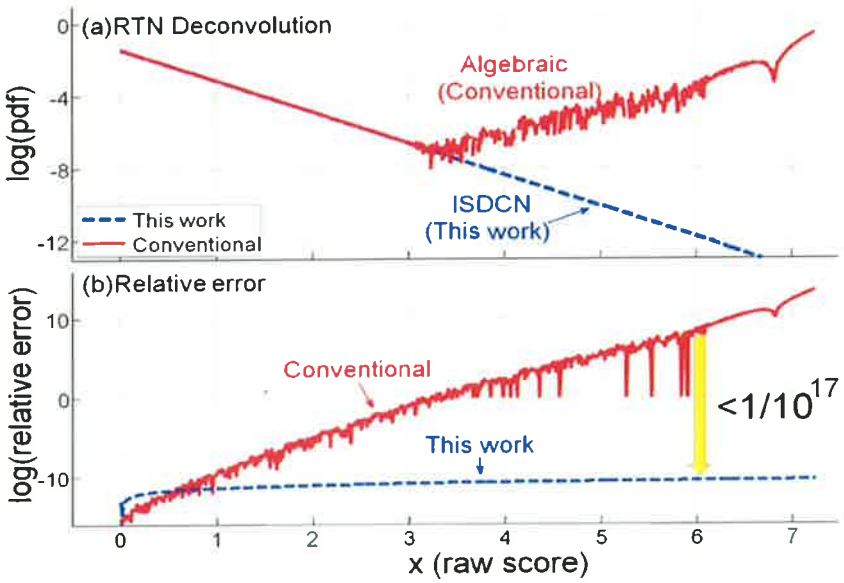


Figure 6.4: The proposed ISDCN successfully eliminates the V-shaped errors by suppressing the accumulated errors.

6.3 Proposed Segmented Deconvolution Algorithm

As shown in Figure 6.3(c), the proposed algorithm takes some optimization process to find the best $f^{(i)}$ that minimize the $|h-h^{(i)}|$. In this application, however, since the distribution type is sort of exponential, the difference in the pdf level between x=0 and x=6 is very huge, e.g., several 10s orders of magnitude, as can be seen in Figure 6.5(a). The conventional

optimization problem prioritizes the higher frequent probability around x=0 because the pdf around there dominates the overall cdf. As a result, unfortunately, the cdf error is best reduced around x=0, while leaving the errors in the tail region (x=6) as it is.

When strictly comparing the relative errors across all x-regions, it has an x-position dependency and becomes larger as the raw score x is increased, as shown in Figure 6.5(c). When considering an actual application for the fail-bit count (FBC) prediction, the degree of precision for the tail distribution gives a huge impact on the accuracy of the FBC prediction because the SRAM fail probability is extremely rare-event level (pdf<10^{12}).

To address this issue, we propose a novel algorithm featuring an iterative and segmented forward problem based deconvolution (ISDCN). The key concept is shown in Figure 6.5(b). The process of seeking the best $f^{(i)}$ follows the sequentially step by step manner, i.e., starting from around x=0 and finally focusing on the important zone of the rare-event so that each optimization step can't be interfered with by the other higher sensitivity zone. Once found the best $f^{(i)}$ in each segmentation, its value is temporally fixed until the next iteration process.

By this algorithm, the error is reduced by 10^3 compared with the non-segmentation case, as shown in Figure 6.5(c). It can be seen in Figure 6.5(c) that the each optimization in individual segmentation is successfully done, resulting in no dependency of the position of the segmentation.

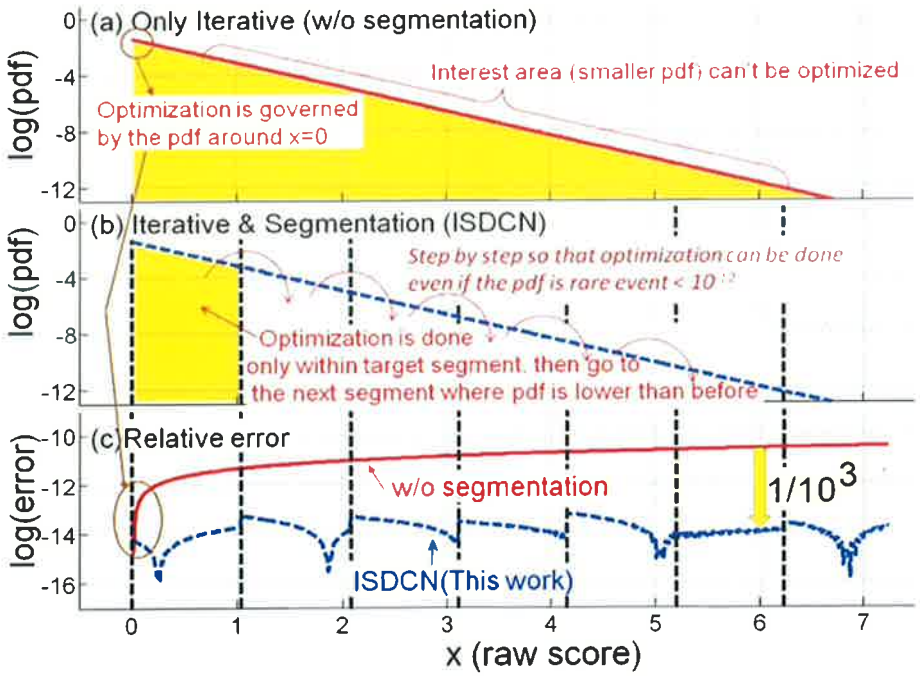


Figure 6.5: Deconvolution errors comparisons between w/ and w/o proposed segmented optimization algorithm.

6.4 RTN Distribution Shape Dependencies

In this session, the RTN tail length (tail slope) dependencies on the error are compared. As can be seen in Figures 6.6(b) and 6.6(c), it is found that the proposed ISDCN can reduce the error for the short tailed and long tailed RTN by 10^2 and 10^4, respectively, compared with the case without any segmented optimizations.

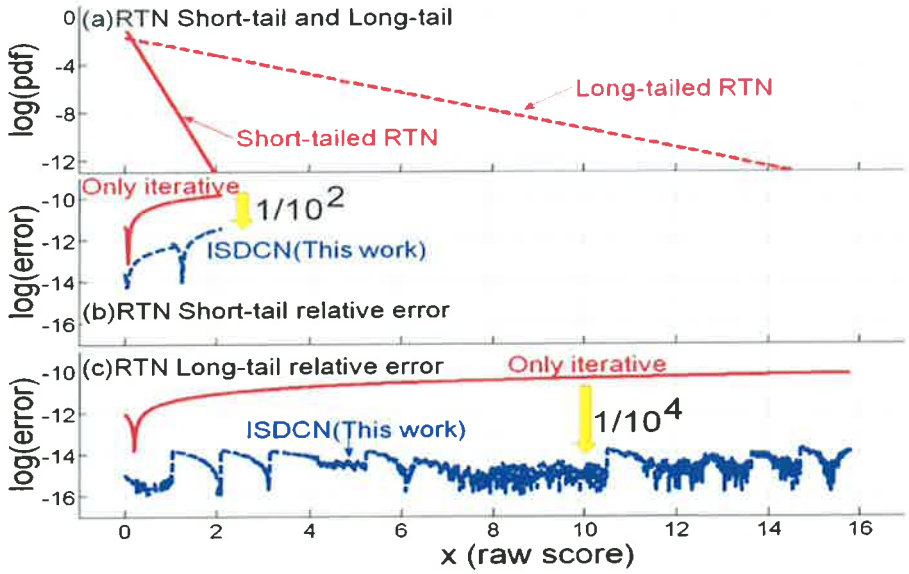


Figure 6.6: Comparisons of the effectiveness of the proposed segmented algorithm between the short and long tailed RTN.

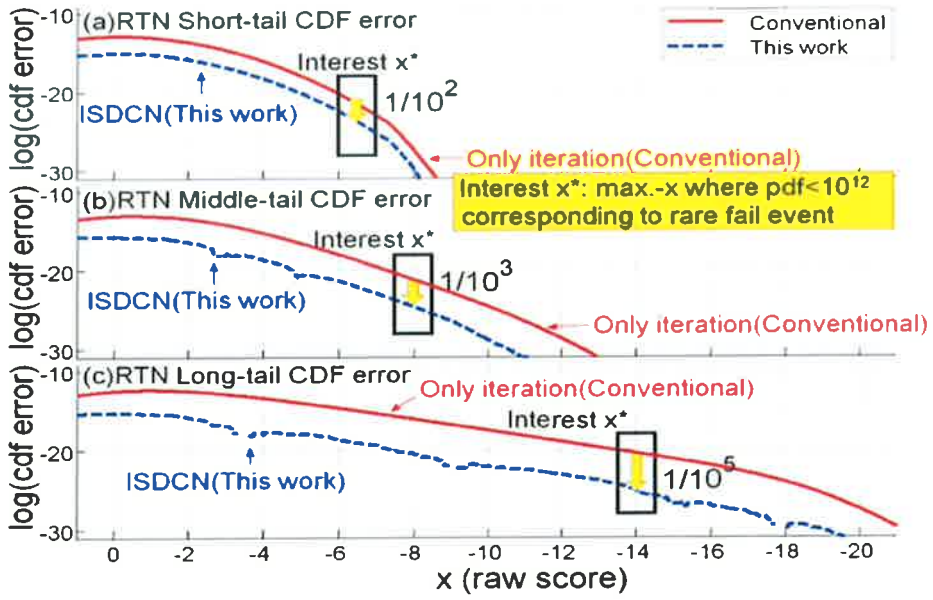


Figure 6.7: Comparisons of cdf error between the cases w/ and w/o the proposed ISDCN algorithm for different RTN tail length.

Since the final target of this work is to increase the accuracy of the FBC prediction, the precision of the cumulative distribution function (cdf) is the key to the success of this research. Figure 6.7 shows the comparisons of the cdf errors for the several cases with a different RTN tail length. The box represents for the interest x-zone where its pdf level is < 10^{12}, i.e., rare event level. This zone is referred to as "interest x". When compared with the cdf error at this x point, it is found that the proposed ISDCN can increase the accuracy of the FBC prediction by 10^2, 10^3, and 10^5, respectively.

6.5 Complex Gamma Mixtures RTN Distributions

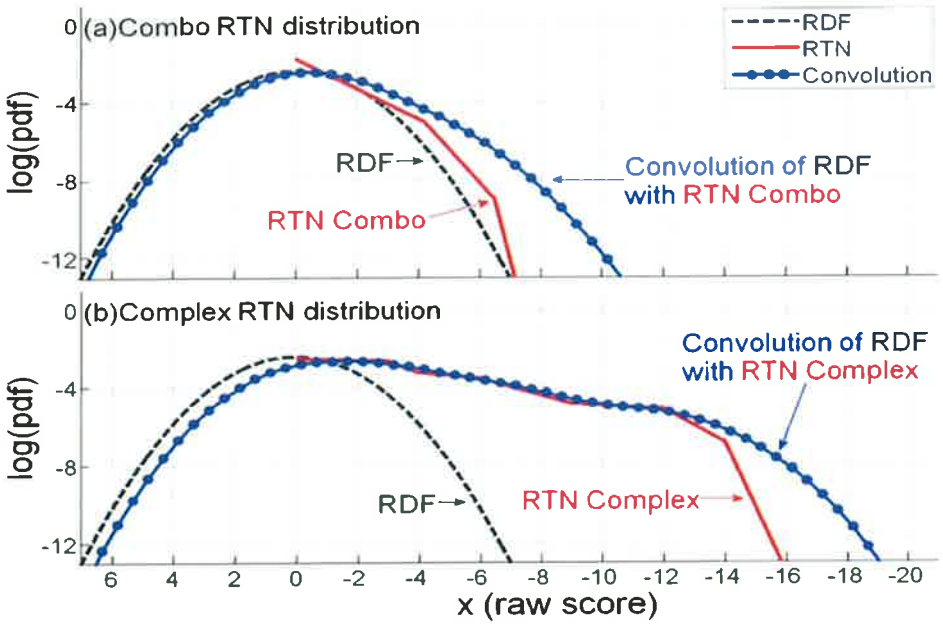


Figure 6.8: 2-types of RTN with (a) multiple convex folding points and (b) mixtures of the convex and concave folding points.

In session 4, assume that the RTN distribution obeys a single gamma distribution with different tail slope and length.

However, we need to assume RTN distribution with more complex gamma mixtures with various sloped segments, as shown in Figure 6.8. In this chapter, refer to the RTN distribution with the multiple convex folding points and the mixtures of the convex and concave folding points, i.e., (a) "RTN combo" and (b) "RTN complex", respectively, as shown in Figure 6.8.

As can be seen in Figure 6.8, the distribution shape gives a big impact on the distribution of the convolution of RDF with RTN. In that sense, have to guarantee the precision level of the RTN deconvolution whatever its slope and length of the RTN tail is.

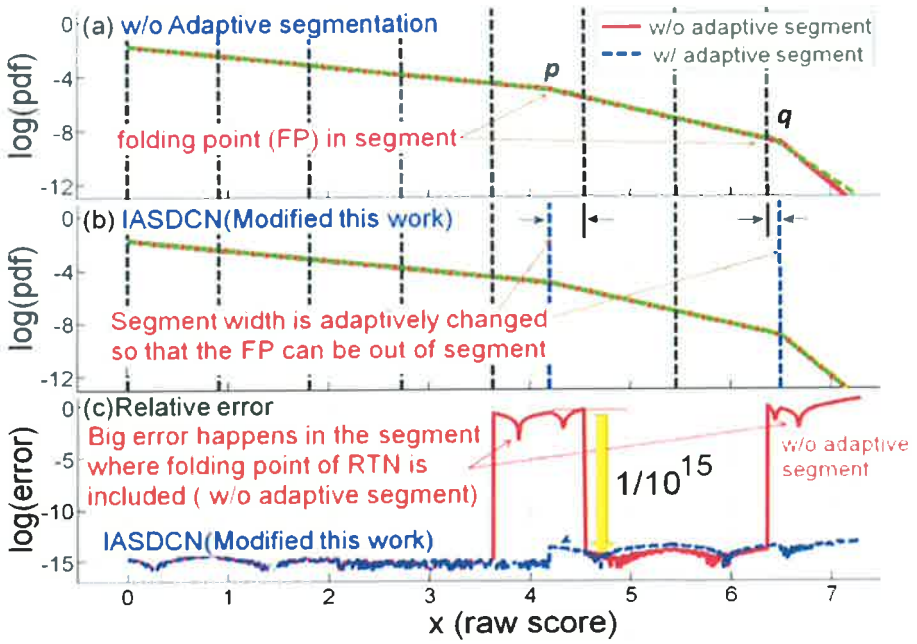


Figure 6.9: Effects of the proposed IASDCN algorithm on the error reduction for the combo-RTN distribution.

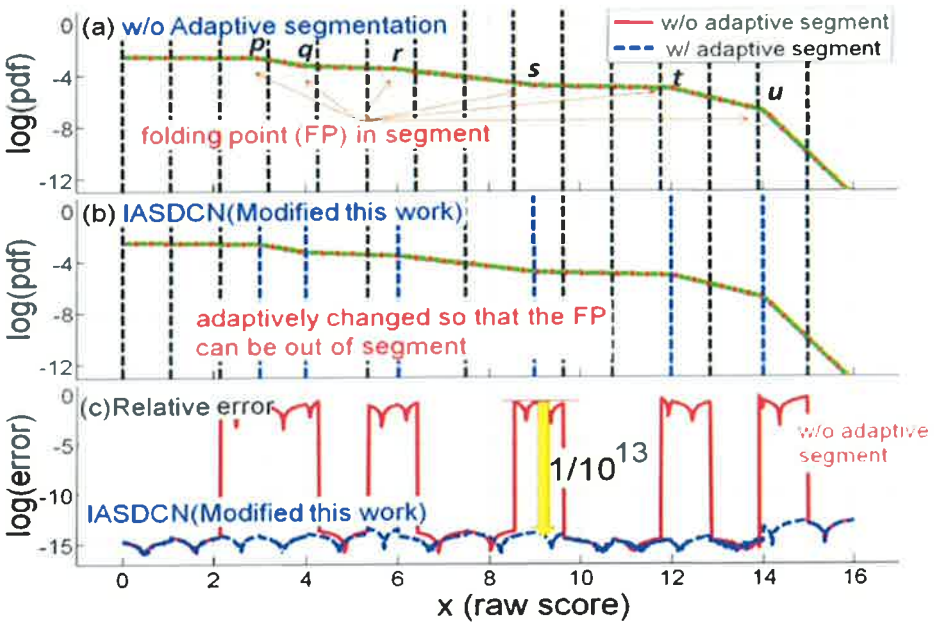


Figure 6.10: Effects of the proposed IASDCN algorithm on the error reduction for the complex-RTN distribution.

To make more clear the effectiveness and remaining issues of the proposed algorithm ISDCN, the demonstration results for the case of combo and complex RTN distributions are shown in Figure 6.9 and Figure 6.10, respectively.

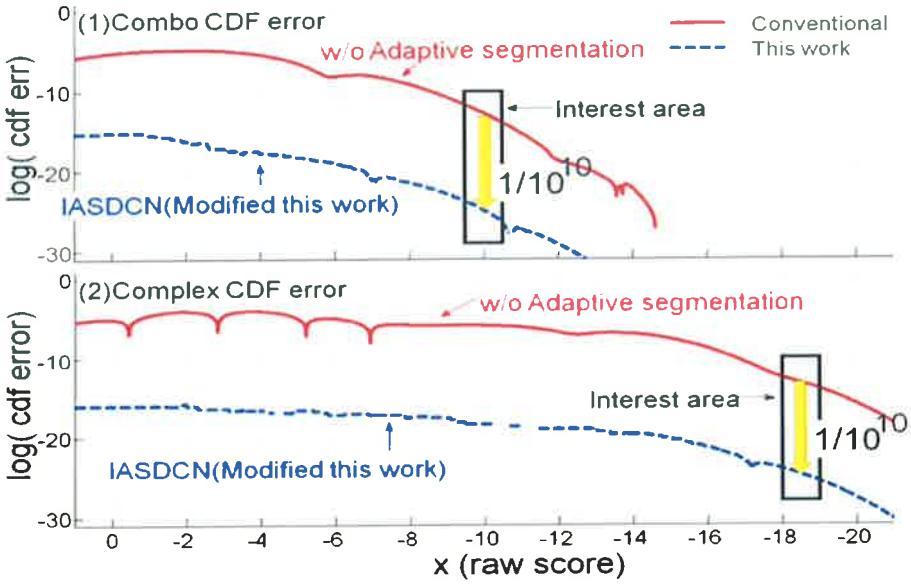


Figure 6.11: Effects of the proposed IASDCN algorithm on the cdf-error reduction for the combo and complex-RTN distribution.

Figure 6.9(a) shows the deconvolution result for the combo RTN with two convex folding points (p,q) by using the ISDCN algorithm. As can be seen in Figure 6.9(c), the errors in the segment that includes the folding points (p,q) are 15-orders of magnitude larger than those in other segments. This is the remaining issue confronting the proposed ISDCN when adapting to the combo and complex typed RTN tail distributions. To address this issue, we propose a new idea for adapting to the original ISDCN. By doing this, the segmentation width can be adaptively changed so that the folding point can be always just on the boundary line of the segmentations, as shown in Figure 6.9(b). By the proposed flexible segmentation, the error is reduced to $1/10^{15}$ of that of original result, as shown in Figure 6.9. We refer this modified work as "iterative and adaptively segmented forward-problem based deconvolution" (IASDCN).

Figures 6.10(a) and 6.10(b) show the comparison of the deconvolution results for the complex RTN with 4-convex (p,r,t,u) and 2-concave (q,s) folding points between the two cases of using the original ISDCN and the modified algorithm (IASDCN), respectively. It is found that the error for IASDCN is reduced to $1/10^{13}$ that of ISDCN, as shown in Figure 6.10(c).

Figure 6.11 shows the comparisons of the cdf error (i.e., FBC prediction error) between the combo and complex RTN with the different RTN tail shape and length. When compared with cdf error at the x point in the interest area, it is found that the modified proposed work (IASDCN) can increase the accuracy of the FBC prediction by 10^{10} compared with the conventional case that doesn't use the adaptive segmentation.

6.6 Conclusions

This chapter has proposed the deconvolution technique (IASDCN) enabling to successfully circumvent the issue on an abnormal V-shaped error confronting the conventional algebraic based-inverse computation process. The effectiveness of the IASDCN algorithm has been demonstrated for the first time with applying to a real analysis for the effects of the RTN and the RDF on the overall SRAM margin variations.

This chapter has shown that the proposed IASDCN technique can reduce its relative RTN deconvolution errors by 10^{13} to 10^{15} times and as a result, the fail-bit-count prediction error can be reduced by 10^{10} thanks to the suppressed deconvolution errors.

6.7 References

[1]. K. Takeuchi, T. Nagumo and T. Hase "Comprehensive SRAM Design Methodology for RTN Reliability", Symposium on VLSI Technology, pp. 130-131, June 2011.

[2]. K. Takeuchi, T. Nagumo, K. Takeda, S. Asayama, S. Yokogawa, K. Imai, Y. Hayashi, "Direct Observation of RTN-induced SRAM Failure by Accelerated Testing and Its Application to Product Reliability Assessment " , in Digest of IEEE Symposium on VLSI Technology 2010, pp. 189-190, 2010.

[3]. M. Yamaoka, H. Miki, A. Bansal, S. Wu, D. J. Frank, E. Leobandung, K. Torii, "Evaluation Methodology for Random Telegraph Noise Effects in SRAM Arrays", Digest of IEEE IEDM, pp. 745-748, 2011.

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

[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]. Q. Shi, O. Kanoun, "Application of Iterative Deconvolution for Wire Fault Location via Reflectometry", Digest of IEEE International Symposium on Instrumentation & Measurement, Sensor Network and Automation (IMSNA), pp. 102-106, 2012.

[7]. W. Somha, H.Yamauchi, "Convolution/ Deconvolution SRAM Analyses for Complex Gamma Mixtures RTN Distributions", Digest of ICICDT-2013, pp. 33-36, 2013.

[8]. W. Somha, H.Yamauchi, "A V-shaped Deconvolution Error Suppression Technique for SRAM Analyses of Long-Tail Gamma Mixtures Random Telegraph Noise Distribution Effects", Digest of ITC-CSCC-2013, pp. 297-299, 2013.

Chapter 7

Blind Deconvolution for Extracting Unknown Two Factors of RTN and Truncated RDF

In this chapter, we propose a blind deconvolution technique for extracting the unknown two variation factors of the RTN and the truncated RDF solely from the given target for SRAM margin variations. Unlike the non-blind deconvolution, the blind deconvolution has to extract the both of the two unknown factors of RTN and truncated RDF simultaneously, that can be sort of ill-posed problem. The proposed algorithm features a sequentially-dual iteration loop and an adaptively segmented forward-problem based blind deconvolution (DIAS-BDCV) process. This allows a free of convergence error in the optimization process. This effectiveness has been demonstrated with applying to a real SRAM design analysis. It has been shown that the proposed DIAS-BDCV technique allows: 1) a free of convergence-error and local-minimum-error in blind deconvolution even if the total number of parameters to be sought in the optimization problem exceeds 20, and 2) a low enough blind deconvolution errors of the RTN and RDF comparable to the level ($< 10^{-13}$) of the non-blind one.

7.1 Introduction

The GB design for the SRAM will become an unprecedentedly crucial challenge because the increased time-dependent (TD) margin variations (MV)-caused failures can't be predicated any more by only the convolution analyses [1-5]. This stems from the facts that unknown MV after the shipment to the market will become much larger than that for the given MV based on the measurements in advance. This indicates that TD-MV is no longer neglected and much attention has to be paid. This leads to an increased pressure to figure out the unknown factors of TD-MV by solving the inverse problem [6-7], although the SRAM designers are unfamiliar with such kind of methodology until now. Thus, this study recently proposed and discussed the "non-blind" deconvolution techniques for the RTN and RDF effects on the SRAM MV [6-8]. However, those assume that either RTN-**g** or RDF-**f** is given in advance when doing the deconvolution of the RDF-**f** or RTN-**g** with the overall

SRAM margin variations, as shown in Figure 7.1(a). This type of deconvolution is referred to as "non-blind" deconvolution. On the other hand, need to assume in some cases that neither RTN-**g** nor RDF-**f** is given in advance, as shown in Figure 7.1(b). This is referred to as "blind" deconvolution [9], but the actual application to the SRAM analyses has not been reported so far to the best our knowledge. Figure 7.1(b) recounts the following scenarios: the target specs are solely predefined. The others are not given in advance. That is to say, both of the RTN distribution **g** and the required truncating point (TP) of the RDF distributions **f** are unknown. The TP should be decided so that the minimum operating voltage (VDDmin) can be guaranteed to avoid any out of specs after the shipment to the market. This has to be solved by the "blind de-convolution" analysis.

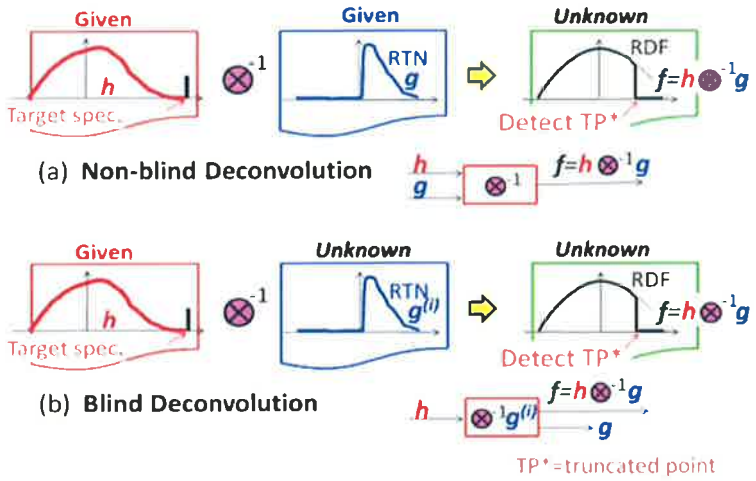


Figure 7.1: Two types of deconvolution: (a) "Non-blind" and (b) "Blind". "Blind" assumes the both RDF & RTN are unknown.

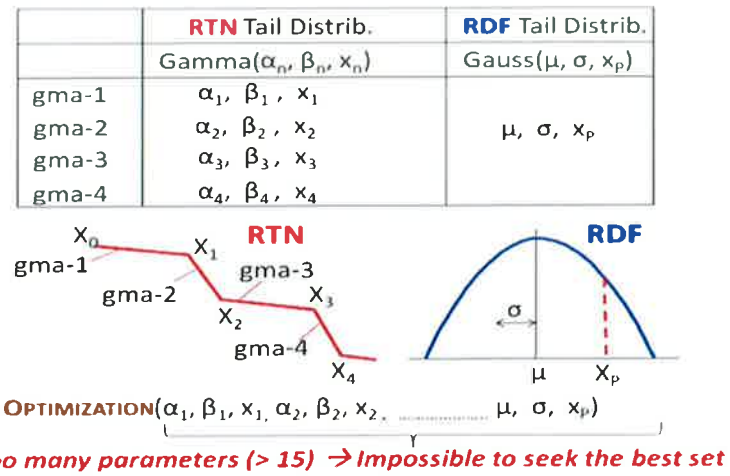


Figure 7.2: The required number of parameters to express the tail distributions for RTN and RDF with Gamma and Gauss.

Compared with the "non-blind" de-convolution, the "blind" deconvolution needs to solve more serious ill-posed problem. For example, the issues on convergence and local minimum errors have to be addressed. In particularly, the increased number of the parameters for the optimization problem makes the blind deconvolution more difficult, as shown in Figure 7.2. The reasons for this trend are: 1) the extended RTN tail length and the increased number of the key factors affecting its slope and length and 2) the increased unknown parameter such as the truncating point TP of the RDF distribution, as shown in Figure 7.2. The purpose of this work is to propose the "blind" deconvolution methods for the SRAM-designs. Compared with the "non-blind" deconvolution, unknown parameters are increased in the "blind" case. Thus, the key to success is to reduce the number of parameters to be sought with fitting algorithms.

To the best of our knowledge, this is the first time to present the blind-deconvolution algorithm for the SRAM-designs featuring a sequentially-dual iteration loop and an adaptively segmented forward-problem based blind-deconvolution (DIAS-BDCV) process and enabling to achieve the above mentioned objects.

7.2 Proposed Iterative Deconvolution Algorithm

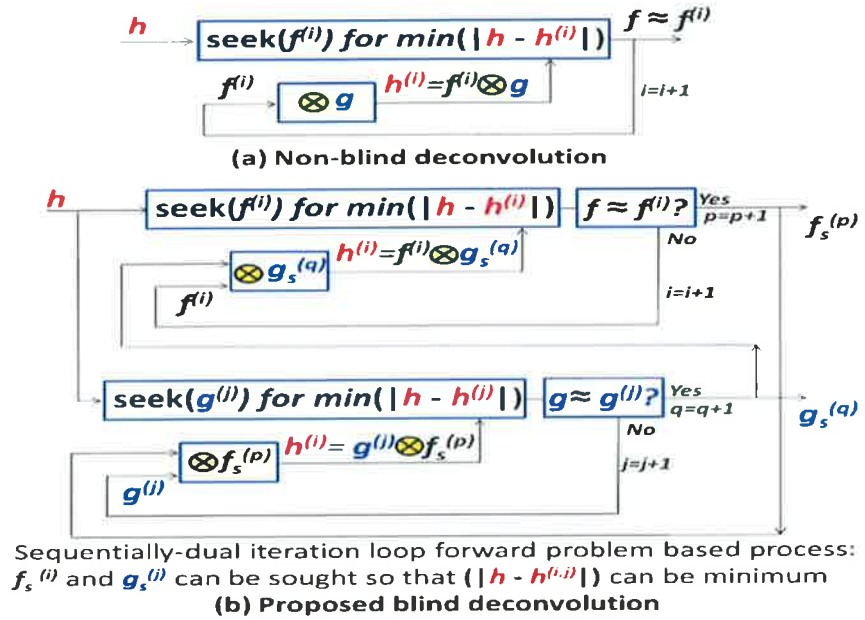


Figure 7.3: Deconvolution algorithm comparisons between (a) non-blind and (b) proposed blind deconvolutions.

The key features of the proposed blind-deconvolution algorithm are: 1) not to use the inverse operation such as the differentiation and division by zero at all. Instead, solve the optimization problem that seeks $\mathbf{f}^{(i)}$ for minimizing ($|\mathbf{h}-\mathbf{h}^{(i)}|$), where $\mathbf{h}^{(i)}$ is the convolution of

$\mathbf{f}^{(i)}$ with g, as shown in Figure 7.3, 2) sequentially-dual iteration loop optimization problem for seeking the parameters separately divided into RTN and RDF. This alleviates the issues on the convergence of the optimization problem thanks to reduced the number of the parameters ($\boldsymbol{\alpha}$,$\boldsymbol{\beta}$,$\mathbf{T_P}$) to be sought, and 3) adaptively segmented optimization to reduce the number of parameters to be sought at once, as shown in Figure 7.4. Since the RTN distribution follows the Gamma mixtures [1-3], each segment can be expressed by the Gamma distribution with 3-parameters of shape $\boldsymbol{\alpha}$, rate $\boldsymbol{\beta}$, and the folding point x_i. If it has 4-folding points, the number of segment is 5, thus, the total numbers of the parameters are 15, as shown in Figure 7.4.

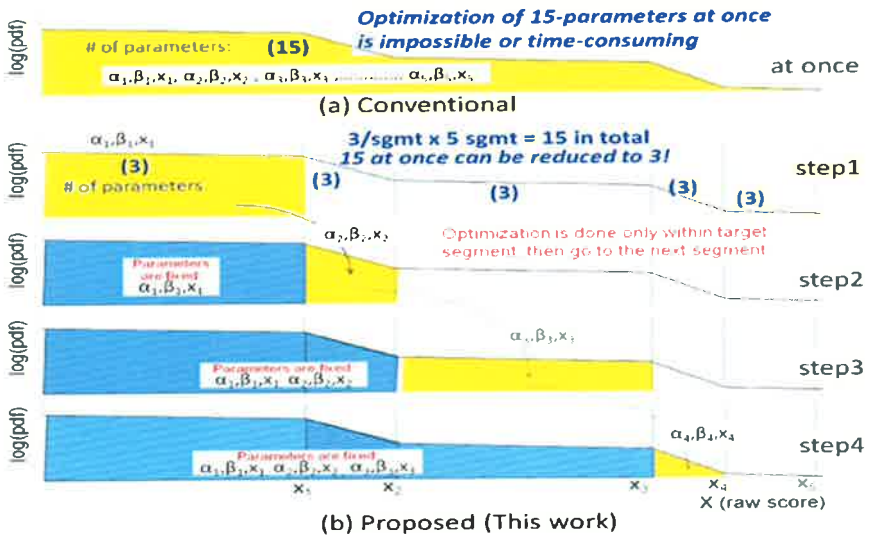


Figure 7.4: Adaptively segmented optimization to reduce the number of parameters to be sought at once.

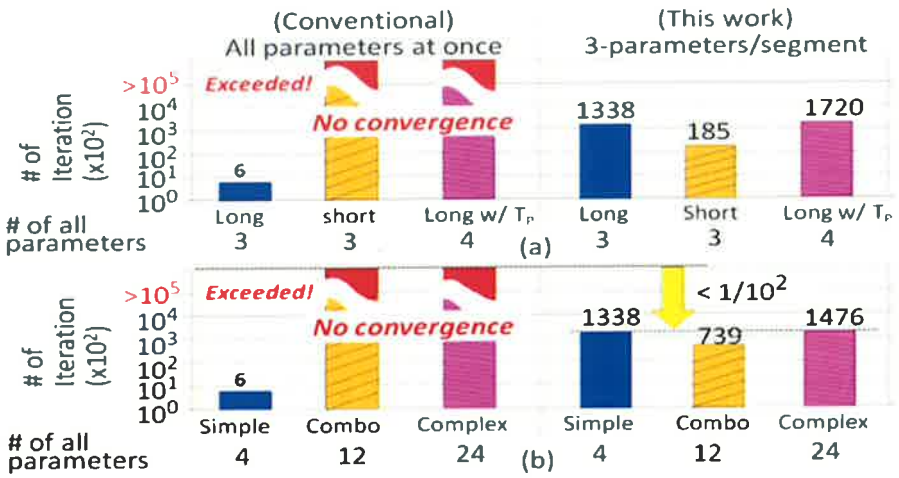


Figure 7.5: Comparisons of the required iteration number of this work and conventional one. (a) shorter and longer than RDF, (b) RTN simple, combo, and complex.

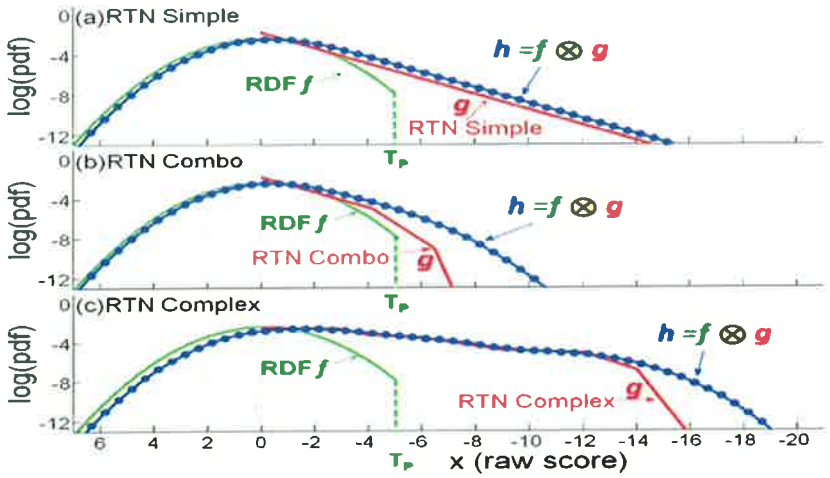


Figure 7.6: (a) RTN simple (long), (b) RTN combo, and (c) RTN complex. This work allows a convergence-error free even for complex RTN.

However, based on our experiments shown in Figure 7.5(a), the number of 3 to 4 is the upper limit for the stable convergence of the optimization. To solve this problem, the adaptively segmented optimization algorithm, in which the parameters to be sought is 3, is proposed. This allows to successfully seek the folding point $\mathbf{x_i}$ as well as $\boldsymbol{\alpha_i}$, and $\boldsymbol{\beta_i}$ (referred in Figure 7.2), resulting in the best determination of all the parameters for the segment, while avoiding any errors in terms of the convergence and the local-minimum. Figure 7.5 shows the comparisons of the required iterations until the convergence between this work and the conventional "at once"-way. The effects of the total number of the parameters on the required iteration numbers are compared for the 3 types of RTN (simple, combo, and complex) and 1 type of truncated RDF (shown in Figure 7.6). Unlike the conventional way, (?)this work allows a stable convergence, which is much less sensitive to the number of all parameters (4, 12, and 24) to be sought. It is also noted that our algorithm provides a less sensitivity of the convergence to the tail length. While the conventional way faces the iteration problem for the case of the simple RTN (short), this work allows a successful convergence for both types of the RTN, as shown in Figure 7.5(a).

7.3 Seeking the Best Parameter Set in Outer Loop

As explained with Figure 7.3, the outer loop of the proposed dual iterative loop in the proposed (DIAS-BDCV) algorithm is to find the three parameters for defining the RDF distribution **f** : (1) truncated point $\mathbf{T_P}$ (referred in Figure 7.1 and 6), (2) standard deviation $\boldsymbol{\sigma}$, and (3) mean $\boldsymbol{\mu}$. Figure 7.7 shows which combination set of the 3-parameters ($\mathbf{T_P}$, $\boldsymbol{\sigma}$, $\boldsymbol{\mu}$) provides the minimum cdf error of the convolution of RTN with RDF, where the RTN distribution **g** is given by the deconvolution results in the inner loop. As shown in Figure

7.7, the cdf error becomes minimum at (T_P =-5 and σ=1). Since there are no other comparable minimum points, the local-minimum issue is not needed to be taken care in this case. Following the case of the RTN simple shown in Figure 7.7(a), the other two cases for the RTN (b) combo and (c) complex are shown in Figures 7.7(b) and 7.7(c), respectively.

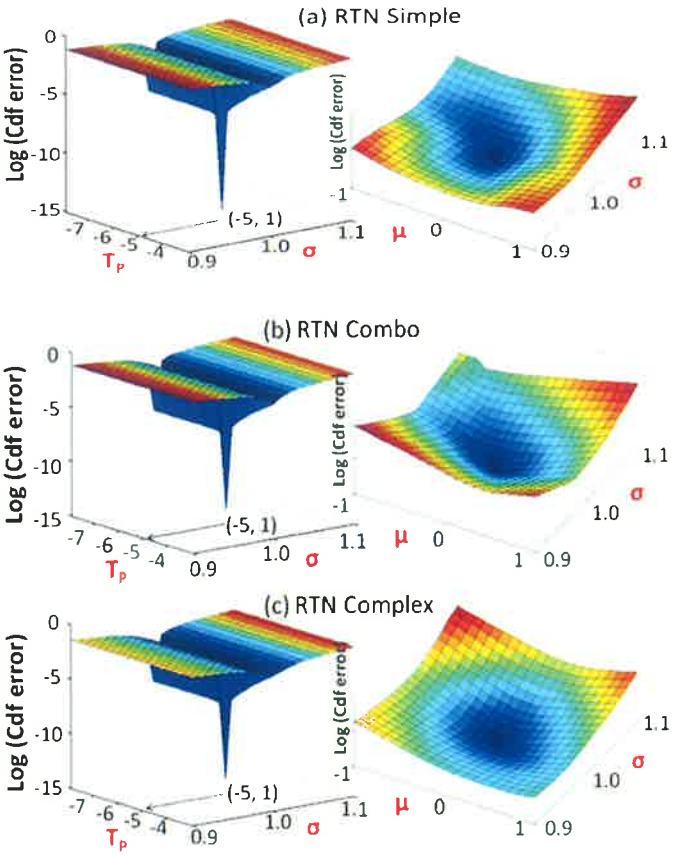


Figure 7.7: Comparisons of seeking the best set in the outer loop for the RTN cases of (a) simple, (b) combo, and (c) complex.

7.4 Comparisons of the Errors between Blind and Non-blind

To verify the effectiveness of the proposed algorithm, we discussed on "how much can the convergence stability in optimization problem be improved?" in the previous sessions 2 and 3. In this session, the level of the accuracy of the deconvolution will be discussed based on the comparisons between the "blind" and "non-blind". In Figure 7.8, "Non-blind RDF" is referred to as the cases that the RTN **g** and the convolution **h=f*g** are given in advance [8]. In the similar way, "Non-blind RTN" is referred to as the cases that the RDF **f** and the convolution **h=f*g** are given in advance. Thus, the required number of the parameters to be sought in the optimization problem for the "Non-blind" is much smaller than that for "Blind". Because of this, the accuracies for "Non-blind" must be much better.

As can be seen in Figure 7.8(b), the error levels of the RDF and RTN for "Blind" are suppressed to the comparable level to "Non-blind", i.e., less than 10^{-13}[8]. Figure 7.9 shows for the case of the "RTN simple (short)". The differences in error level are much smaller than that for the case of "long".

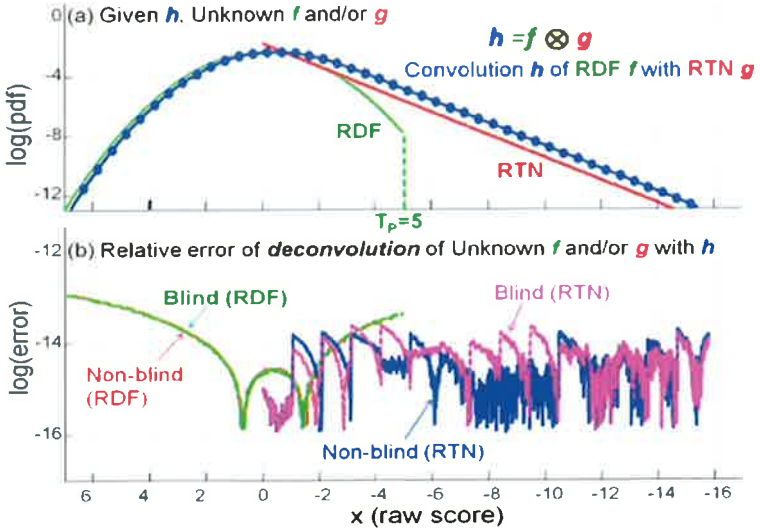


Figure 7.8 Comparisons of the deconvolution errors b/w the cases for blind and non-blind deconvolution. (long RTN simple).

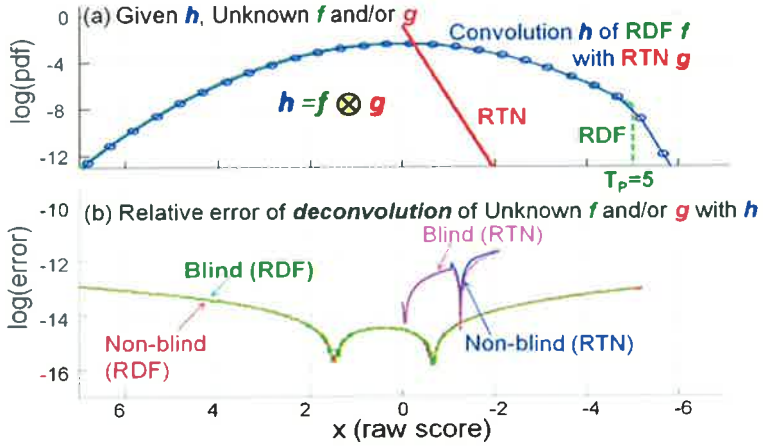


Figure 7.9 Comparisons of the deconvolution errors b/w the cases for blind and non-blind deconvolution. (Short RTN simple).

Figure 7.10 shows for the two cases of (a) the "RTN combo" and the "RTN complex", respectively. The differences in the errors are much smaller than that for the case of "long". The level of the error for "Blind" is comparable to that for "Non-blind" [8]. As can be seen in Figure 7.5(b) and 7.10, the range of the raw score x for the "RTN complex" is wider because of longer tail than that for the "RTN combo". According to the measured iteration

numbers shown in Figure 7.5(b), more iteration cycles are needed for the "longer tails". However, the error level even for "longer tails" can be kept the comparable level to that for "shorter tails".

We conclude that the proposed algorithm makes the errors and iteration numbers less sensitive to the distribution-shape of the RTN/RDF. It can suppress the "Blind-deconvolution" errors to low enough level thanks to the sequentially-dual iteration loop and adaptively segmented forward-problem based process.

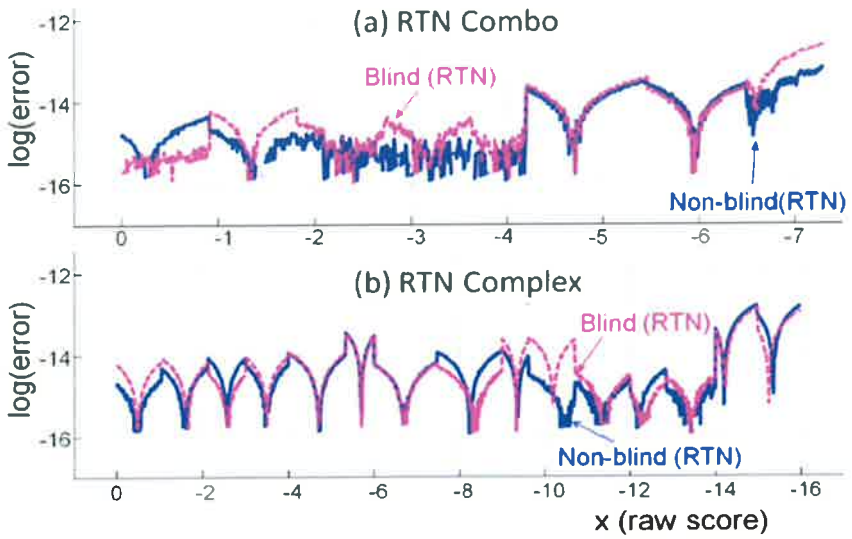


Figure 7.10 Comparisons of the deconvolution errors b/w the cases for blind and non-blind deconvolution for (a) RTN combo and (b) RTN complex.

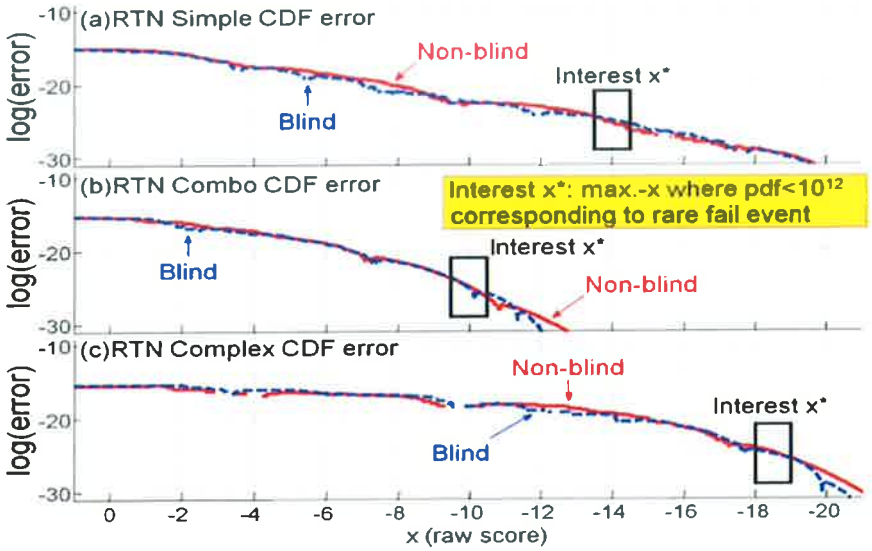


Figure 7.11 Comparisons of the FBC prediction errors (cdf error) for blind and non-blind deconvolution for the RTN (a) simple, (b) combo and (c) complex.

Figure 7.11 shows for the comparisons of the cumulative–density-function (cdf) error (i.e., FBC prediction error) between the three cases of the shapes of the RTN tail: (a) simple, (b) combo and (c) complex. This cdf for the fail bit count (FBC) prediction is given based on the convolution of RDF with RTN, i.e., **h**=**f**(RDF)*g(RTN).

When compared with the cdf error of **h** at the x point (raw score) in the interest area, it is found that the proposed blind deconvolution (DIAS-BDCV) can guarantee the low enough accuracy level comparable to that of "Non-blind" deconvolution [8]. As can be seen in Figure 7.11, the error level is less than 10^{-25} at the interest x. Since the pdf=10^{-12} at the interest x, a 10^{-25} of this error level is small enough. Even assuming the total SRAM bits density is 1Gbit and its yield is 99.9%, the error level corresponds to 10^{-13}/bit of the FBC error, where it is roughly given from the simple calculation of "10^{12} (bits) x 10^{-25} (error)=10^{-13} (bit)" .

7.5 Conclusions

This chapter has proposed the blind deconvolution technique (DIAS-BDCV) enabling to successfully circumvent the issues on various errors in terms of the convergence and division by zero confronting the conventional computation algorithm. The effectiveness of the DIAS-BDCV algorithm has been demonstrated for the first time with applying to a real analysis for the effects of the RTN and RDF on the overall SRAM margin variations. This chapter has shown that the proposed DIAS-BDCV technique can allow useful blind-deconvolution that is capable of suppressing the FBC prediction error to a 10^{-13}/bit thanks to the suppressed blind deconvolution errors.

7.6 References

[1]. X. Wang, A.R. Broun, B. Cheng, A. Asenov, "RTS amplitude distribution in 20nm SOI FinFETs subject to Statistical Variability", Simulation of Semiconductor Processes and Devices SISPAD 2012, pp.296-299, 2012.

[2]. 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.

[3]. M. Yamaoka, H. Miki, A. Bansal, S. Wu, D. J. Frank, E. Leobandung, K. Torii, "Evaluation Methodology for Random Telegraph Noise Effects in SRAM Arrays", Digest of IEEE IEDM, pp. 745-748, 2011.

[4]. K. Takeuchi, T. Nagumo and T. Hase "Comprehensive SRAM Design Methodology for RTN Reliability", Symposium on VLSI Technology, pp. 130-131, June 2011.

[5]. K. Takeuchi, T. Nagumo, K. Takeda, S. Asayama, S. Yokogawa, K. Imai, Y. Hayashi, "Direct Observation of RTN-induced SRAM Failure by Accelerated Testing and Its Application to Product Reliability Assessment " , in Digest of IEEE Symposium on VLSI Technology 2010, pp. 189-190, 2010.

[6]. W. Somha, H.Yamauchi, "Convolution/ Deconvolution SRAM Analyses for Complex Gamma Mixtures RTN Distributions", Digest of ICICDT-2013, pp. 33-36, 2013.

[7]. W. Somha, H.Yamauchi, "A V-shaped Deconvolution Error Suppression Technique for SRAM Analyses of Long-Tail Gamma Mixtures Random Telegraph Noise Distribution Effects", Digest of ITC-CSCC-2013, pp. 297-299, 2013.

[8]. W. Somha, H.Yamauchi "Iterative and Adaptively Segmented Forward Problem Based Non-Blind Deconvolution Technique for Analyzing SRAM Margin Variation Effects", Digest of ISOCC-2013, pp. T.B.D. (accepted to appear), 2013.

[9]. Fish D.A. A.M. Brinicombe, E.R. Pike, "Blind deconvolution by means of the Richardson–Lucy algorithm", Journal of the Optical Society of America A12 (1), pp.58–65, 1995.

Chapter 8

RTN Deconvolutions with Richardson-Lucy and Proposed Segmentation Algorithms

In this chapter, a comparative analysis of the RTN deconvolution accuracy between the Richardson-Lucy (R-L) algorithm and the proposed partitioned forward-problem-based deconvolution means (PFDCV) is presented. Unlike the R-L based deconvolution, the proposed technique successfully solves the issue of noise amplification thanks to eliminating any operations of differential and division. This effectiveness has been demonstrated by applying it to a real analysis for the effects of the RTN on the overall SRAM margin variations. It has been shown that the proposed PFDCV technique can reduce its relative errors of the RTN deconvolution by 10^{14}–fold compared with the cases of the R-L.

8.1 Introduction

The approximation-error of the tails of RTN distribution will become a crucial challenge. This stems from the facts that: (1) tails of the RTN distribution (**g**) will become longer than that of random-dopant-fluctuation (RDF) (**f**) that is previously dominant factor of overall margin-variations, as shown in Figure 8.1 and (2) the convolution result (**h=f⊗g**) of the RDF(**f**) with the RTN(**g**) will be more governed by the RTN than the RDF, as can be seen in the comparison of (**h=f⊗g**) between Figure 8.2(a) and 8.2(b) for short and long RTN, respectively. Because the increasing paces of variation-amplitude σV_{th} of the threshold voltage (V_{th}) are differently dependent on the MOSFET channel-size (LW) like the below expressions of (1) and (2), the σV_{th} increasing paces of the RTN is a 1.4x faster than that of the RDF if assuming the LW is scaled down by 0.5-fold every process generation, as shown in Figure 8.1(a). According to the references [1]-[4], there will come the time soon around a 15nm-scaled CMOS era.

The reliability design for the SRAM will become an unprecedentedly crucial challenge because the increased time-dependent (TD) margin variations (MV)-caused failures cannot be predicated any more by only ordinary **convolution** analyses [1]-[4].

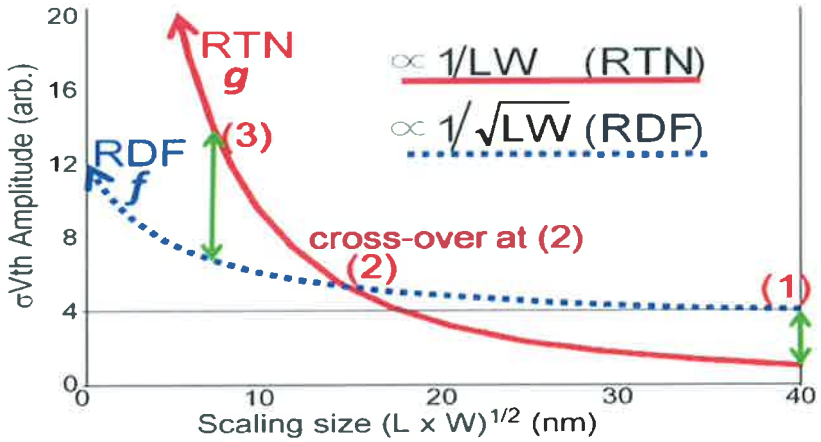


Figure 8.1: (a) Trend of variation amplitude of RTN and RDF. Variation amplitude of RTN becomes larger than that for RDF in 10nm era.

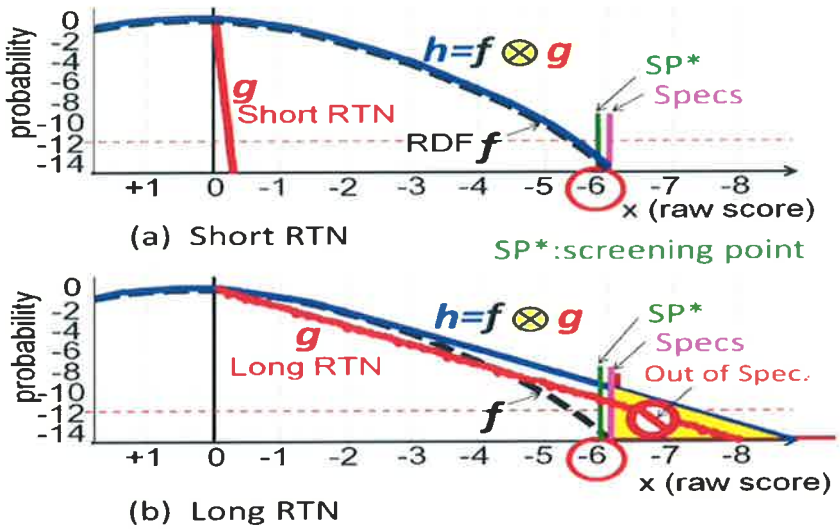


Figure 8.2: Comparison of the convolution result **h=f⊗g** of the RDF(**f**) with RTN(**g**) between (a) short RTN and (b) long RTN.

This stems from the facts that latent TD-MV, (i.e., unknown MV after shipped to the market), will become much larger than already-known MV based on the measurements in advance. This leads to an increased pressure to figure out the unknown factors by solving the inverse problem [5]-[9], although the SRAM designers are unfamiliar with such kind of methodology until now.

Figures 8.3(a) and 8.3(b) show an example for the **deconvolution** ($\otimes^{-1}$) and the **convolution** ($\otimes$), respectively. Where $\otimes^{-1}$ and $\otimes$ are arithmetic symbol for **deconvolution** and convolution, respectively.

Figure 8.3(a) recounts the following scenarios: a certain distribution (**h**) within the product target spec (SP_{prod}) is predefined and the RDF distribution (**f**) is already-known based on the measured data. The **f** is truncated at a certain point (TP) based on the screening spec and converted to $\mathbf{f_{TP}}$. However, the **TP** of the $\mathbf{f_{TP}}$ and the RTN distribution (**g**) are unknown and should be decided as the screening spec and process target spec, respectively, such that the **h** can be within the $\mathbf{SP_{prod}}$, as shown in Figure 8.3(a). The **g** is calculated by the deconvolution of ($\mathbf{g=h\otimes^{-1}f_{TP}}$). Unlike the convolution ($\mathbf{h=f_{TP}\otimes g}$) (Figure 8.3(b)), the deconvolution of the **g** is sort of ill-posed problem and troublesome operation [5]-[8].

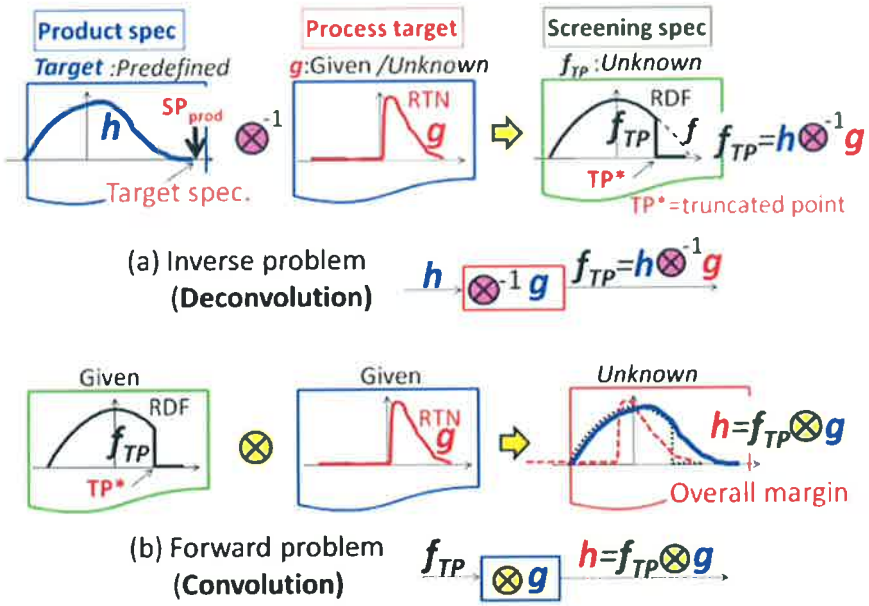


Figure 8.3: (a) inverse problem (deconvolution $\otimes^{-1}$) (b) forward problem (convolution $\otimes$).

8.2 Discussions on the Ricahrdson-Lucy (R-L) Deconvolution of RTN Accuracy

8.2.1 Richardson-Lucy Deconvolution Algorithm

Richardson-Lucy (R-L) algorithm [9] is one of the most widely used deconvolution algorithms (See Figure 8.4) in the area of image processing although it has some shortcomings such as noise amplification [9]. As can be seen in Figure 8.4, the calculation process is based on an iteration and convolution ($\otimes$). However, it relies on the maximum likelihood iterations [9] and needs some derivative operation. As one of the tools for the deconvolution, the "**deconvlucy**" has been built in MATLAB®.

However, to the best of our knowledge, there have been no examples of the R-L deconvolution being used for the SRAM margin analyses, to demonstrate the issues of the R-L deconvolution of the RTN distribution (**g**) with the SRAM MV (**h**). Figure 8.5(a) shows the issue of an abnormal ringing error (noise amplification) confronting the R-L algorithm based deconvolution. This is due to an unstable division operation and maximum likelihood iterations [9]. Even if adjusting the damping factor [10]-[11], severe side effect from the damping is unfortunately caused around x=0 to -4 (see Figure 8.5(b)).

As a result, the deconvoluted RTN distribution is significantly deviated from the expected curve (see Figure 8.5).

Unlike the application in the area of image processing, "rare-event fail bit count (FBC) analyses for the SRAM design" is very susceptible to the error of the probability density function (pdf) even if the pdf error level is as small as 10^{-12}.

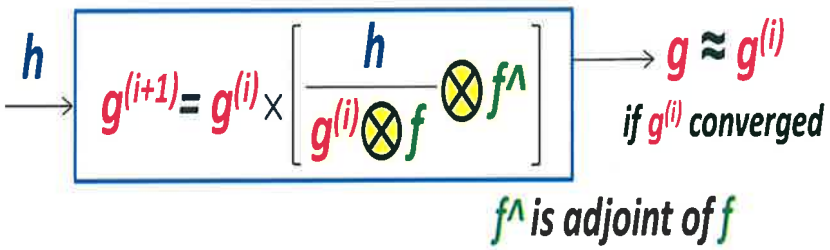


Figure 8.4: (a) Algorithm of Richardson-Lucy deconvolution of RTN(g).

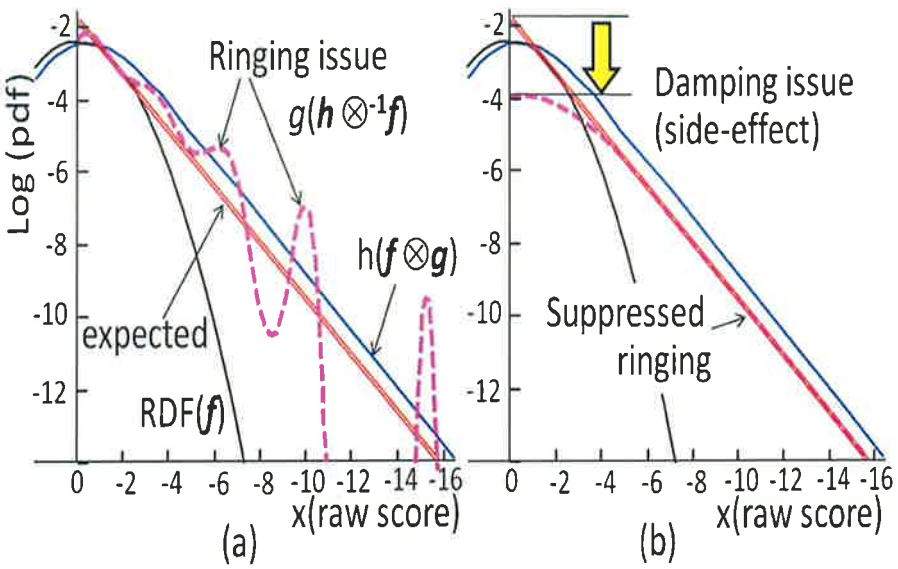


Figure 8.5: R-L deconvolution of RTN(g) (a) ringing occurs (b) ringing is suppressed with damping factor=0.5.

As explained in this subsection, the RTN R-L deconvolution errors have some dependencies on: 1) the number of iteration cycles and 2) damping factors. Thus, in the following subsections, the detailed analyses of the dependencies of the error will be discussed including another dependency of the RTN tail length and shape.

8.2.2 Iteration Cycles Dependency of Deconvolution Errors

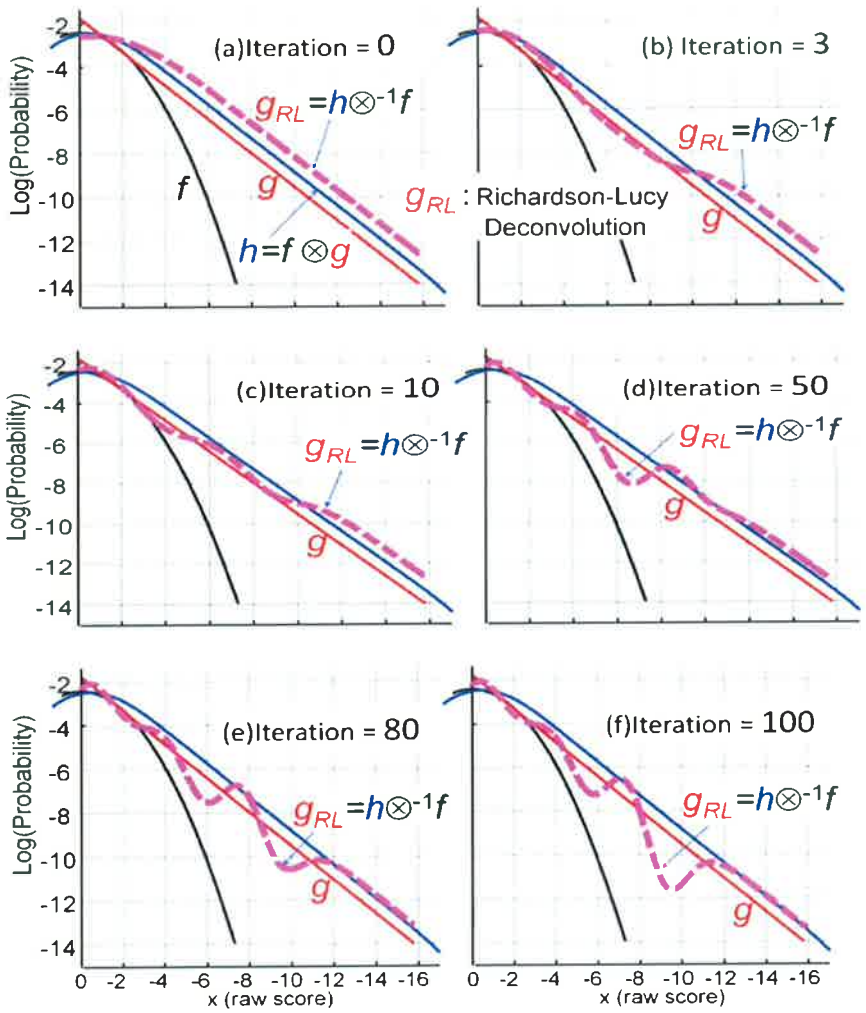


Figure 8.6: Iteration cycle dependencies of the deconvolution, N is iteration numbers (a)N=0, (b)N=3, (c)N=10, (d)N=50, (e)N=80, and (f)N=100.

Figure 8.6 shows the iteration cycle dependencies of the RTN **g** deconvolution with Richardson-Lucy (R-L) algorithm ($\mathbf{g_{RL}=h\otimes^{-1}f}$). Where **h** is the convolution of the RDF(**f**) with the RTN(**g**), i.e., (**h=f⊗g**). Relative R-L deconvolution error (g_{RL_ERROR}) is defined by the following expression of (8.1).

$$g_{RL\_ERROR}=|\mathbf{g(x)}-\mathbf{g_{RL}(x)}|\,/\,\mathbf{g(x)} \qquad \text{--------- (8.1)}$$

As can be seen in Figures 8.6(a)-8.6(f), the RTN R-L deconvolution $\mathbf{g_{RL}}$ has a complex dependency on the iteration cycles. The frequency and amplification of the ringing curve are changed with the iterations. However, it is hard to find the best one based on the iteration dependency because of its complexity.

8.2.3 RTN Tail-length Dependency of Deconvolution Errors

As explained with Figure 8.1, the amplitude of the V_{th} shift caused by the RTN is increased with the scaling trend. The RTN1, RTN2, and RTN3 (see Figures 8.7(a), 8.7(b), and 8.7(c), respectively) correspond to the RTN distributions at the position of (1), (2), and (3) in Figure 8.1, respectively. To make the RTN-length dependencies clearer, the R-L deconvolution of the RTN $\mathbf{g_{RL}}$ were compared among the RTN1, RTN2, and RTN3, while changing the iteration cycles N=10,100, and 1000, as shown in Figure 8.8 and Figure 8.9.

Figures 8.8(a), 8.8(b), and 8.8(c) show the R-L deconvolutions of the RTN1 and the RTN2 at the iteration cycles N= 10, 100, and 1000, respectively.

It is found that no ringing happens in the R-L deconvolution of the RTN1 unlike the case for the RTN2 even if the iteration number N is increased up to 1000. The relative error of the RTN1 deconvolution $\mathbf{g_{RL}}(x)$ has an x-position dependency. As the N is increased, the relative error of $\mathbf{g_{RL}}(x)$ for RTN1 in the region of x=0 to -1 is reduced while the error in the region of x=-1 to -2 is increased.

Figures 8.9(a), 8.9(b), and 8.9(c) show the R-L deconvoltions of the RTN3 at the iteration cycles N= 10, 100, and 1000, respectively.

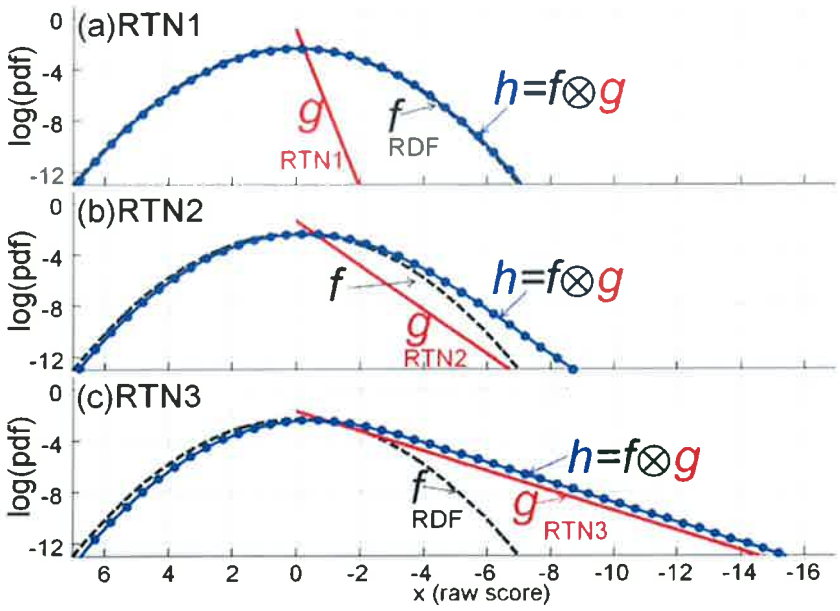


Figure 8.7: Relationship of the RTN tail-length among the RTN1, RTN2 and RTN3.

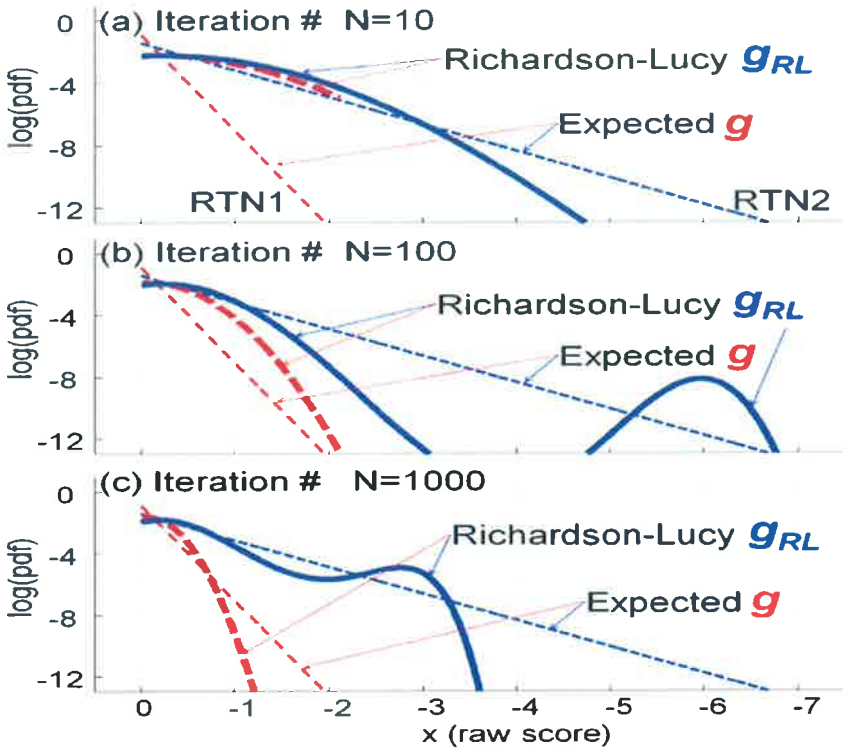


Figure 8.8: Iteration cycle dependencies of the deconvolution for RTN1 and RTN2 for (a) N=10, (b)N=100, and (c)N=1000.

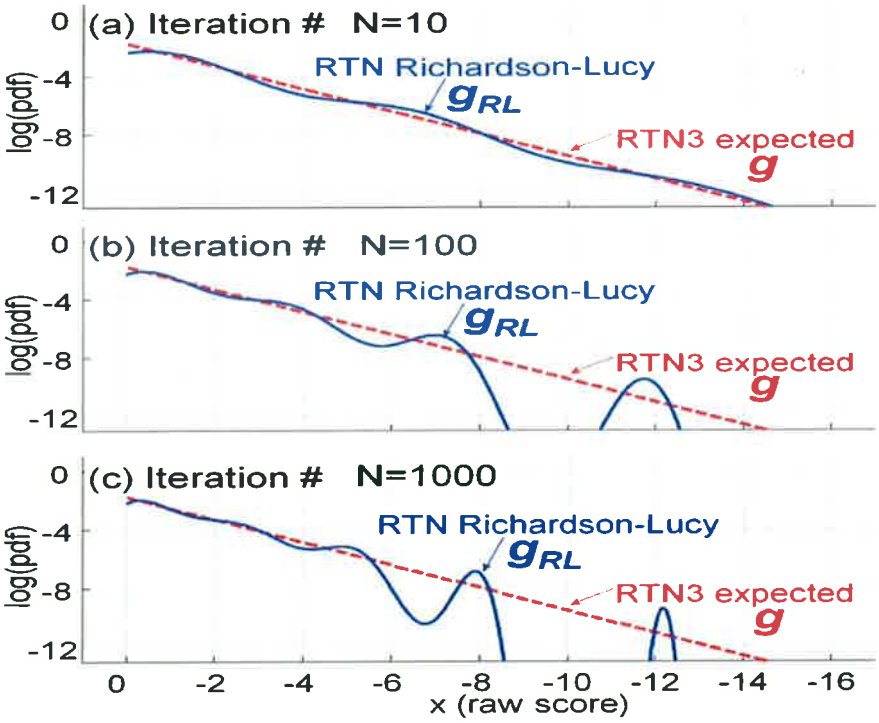


Figure 8.9: Iteration cycle dependencies of the deconvolution for RTN3 for (a) N=10, (b) N=100 and (c) N=1000.

It is found that significant ringing happens in the R-L deconvolution of the RTN3 unlike the case for the RTN1 if the iteration number N is increased up to 1000. The relative error of the RTN3 deconvolution $\mathbf{g_{RL}}(x)$ has a strong x-position dependency. As the N is increased, the relative error of $\mathbf{g_{RL}}(x)$ for the RTN3 in the region of x=0 to -2 is reduced while the error in the region of x=-4 to -12 is increased due to the ringing and its amplification.

8.2.4 RTN Shape Dependency of Deconvolution Errors

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 Figure 8.10. 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 Figure 8.10(a). In this study, this shape of the RTN distribution is referred to as "Combo". Figure 8.10(b) shows the more complex shape comprising of the multiple line-segment of (O-P), (P-Q), (Q-R), (R-S), (S-T), (T-U) and (U-V) with different slope. The multiple line-segments are connected at the concave and convex folding points of O, P, Q, R, S, T, U and V, as shown in Figure 8.10(b).

Figures 8.11(a), 8.11(b), and 8.11(c) show the R-L deconvolutions of the "Combo" at the iteration cycles N of 10, 100, and 1000, respectively.

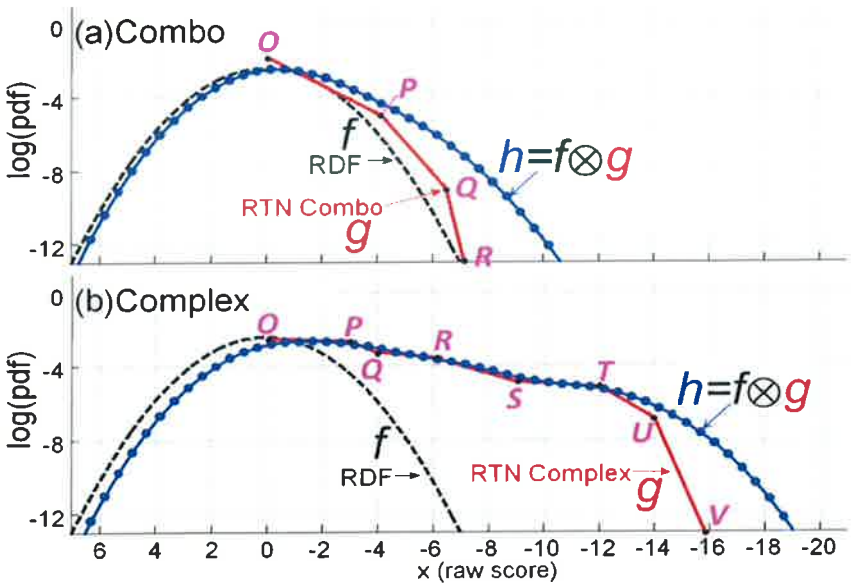


Figure 8.10: Relationship of the RTN tail-shape between "Combo" and "Complex"

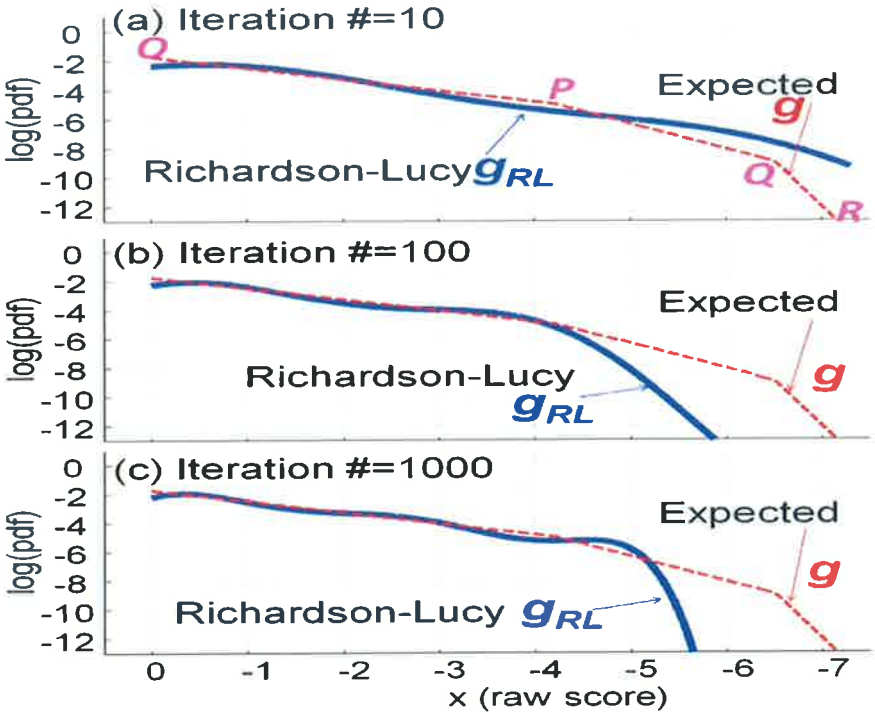


Figure 8.11: Iteration cycle dependencies of the deconvolution for "Combo" for (a) N=10, (b) N=100 and (c) N=1000.

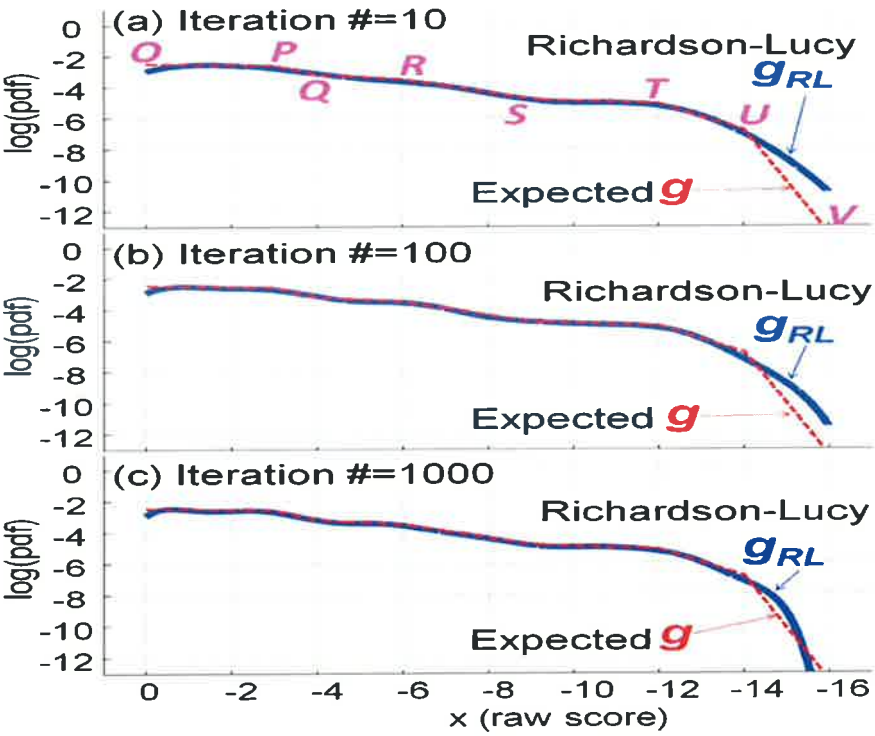


Figure 8.12: Iteration cycle dependencies of the deconvolution for "Complex" for (a) N=10, (b) N=100 and (c) N=1000.

It is found that the relative error of the "Combo" deconvolution $\mathbf{g_{RL}}(x)$ has an x-position dependency. As the N is increased, the relative error of $\mathbf{g_{RL}}(x)$ in the region of x=0 to -3 is reduced while the error in the region of x=-3 to -7 is increased.

Figures 8.12(a), 8.12(b), and 8.12(c) show the R-L deconvolutions of the "Complex" at the iteration cycles N of 10, 100, and 1000, respectively.

It is found that the relative error of the "Complex" deconvolution $\mathbf{g_{RL}}(x)$ has an x-position dependency. As the N is increased, the relative error of $\mathbf{g_{RL}}(x)$ in the region of x=0 to -12 is reduced while the error in the region of x=-12 to -16 is also reduced unlike the cases of the RTN2, RTN3, and "Combo".

8.2.5 Cycle and RTN Tail Dependency of Deconvolution Errors

Convergence properties of the Richardson-Lucy iteration process for the deconvolution for the different tails of the RTN1, RTN2, RTN3, Combo, and Complex are compared, as shown in Figure 8.13 and Figure 8.14, respectively.

It is found that the convergence properties for the RTN2 and the "Combo" are not secured. Error amplitude of the cumulative density function (cdf) are not converged but oscillated when the iteration number N is increased up to 10^5, as shown in Figures 8.13 and 8.14. The convergence behavior is similar between the RTN2 and "Combo". The common factor of the two is the length of the tail, i.e., average gradient of the slope, as can be seen in Figures 8.7(a) and 8.10(b).

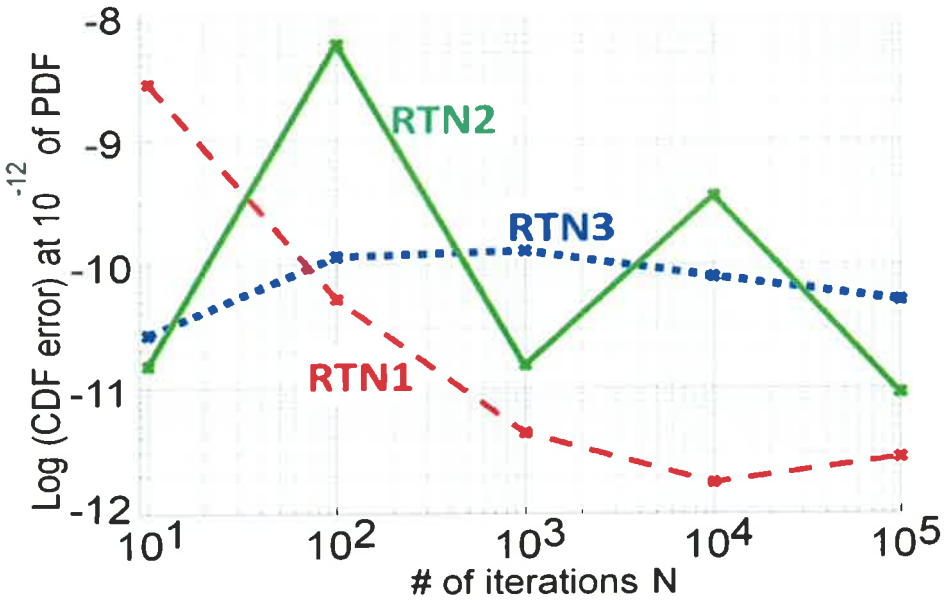


Figure 8.13: Iteration cycle dependencies of the deconvolution for "Complex" for (a) N=10, (b) N=100 and (c) N=1000.

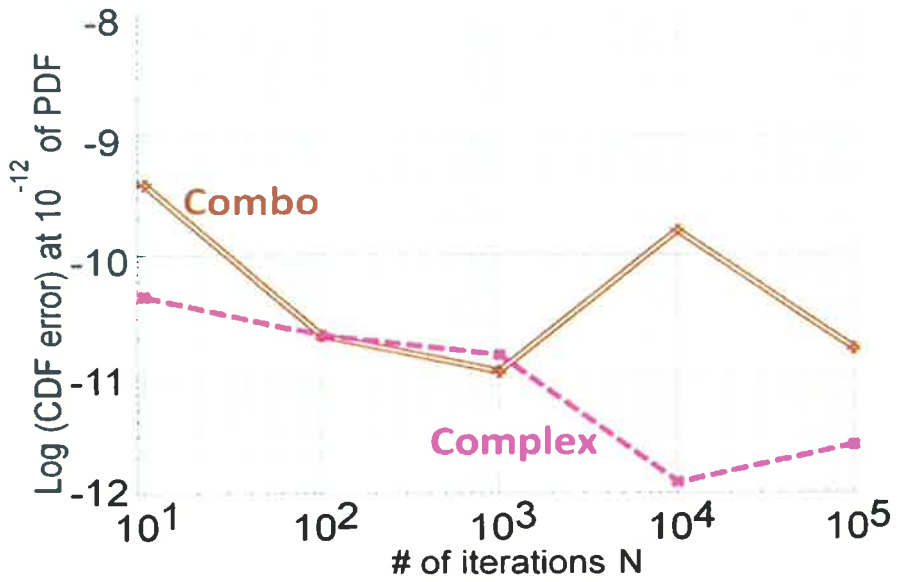


Figure 8.14: Iteration number N dependencies of the deconvolution for "Combo" and "Complex".

The error of the cdf is defined as the following expression (8.2).

$$cdf_{ERROR}=|cdf_{RL}(-Xp)-cdf(-Xp)|/cdf(-Xp) \quad \text{---------(8.2)}$$

Where, cdf_{RL} is the cdf of deconvoluted RTN by the R-L algorithm. Xp is the point where pdf=10^{-12}

8.2.6 Damping Factor Dependency of Deconvolution Errors

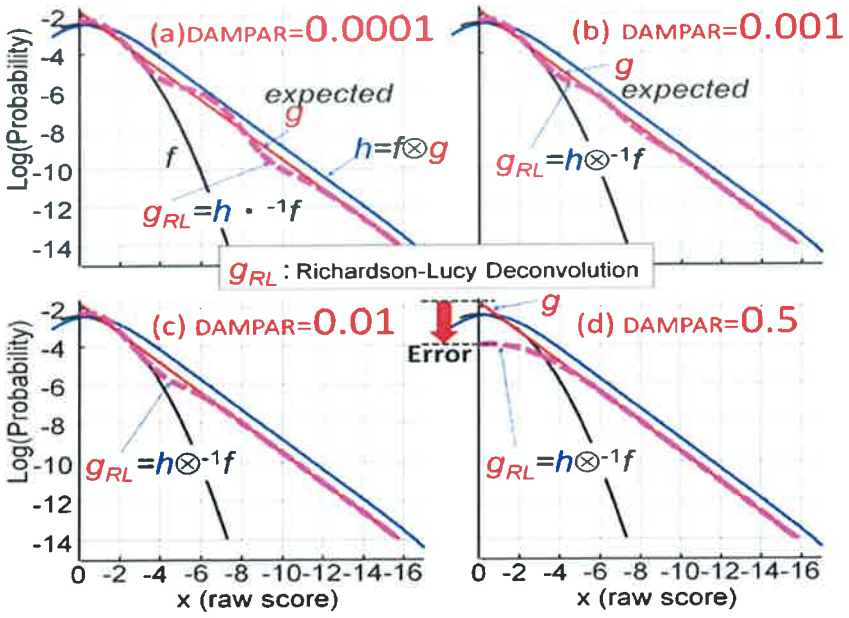


Figure 8.15: Damping factor dependencies of the deconvolution for RTN3.

The built-in function of "**deconvlucy**(DAMPAR)" in MATLAB® can specify the threshold deviation of the resulting image. Iterations are suppressed for pixels that deviate beyond the "DAMPAR" value from their original value.

Even if adjusting the damping factor [10]-[11], severe side effect from the damping is unfortunately caused around x=0 to -4, as shown in Figure 8.15(d). The $\mathbf{g}_{RL}$ is deviated from the expected **g** of the RTN.

8.3 Proposed Partitioned Forward-Problem Based Deconvolution (PFDCV)

Our proposed idea considers the rare event probability area by introducing the segmented optimization. This is different from the conventional optimization problem that unfortunately attempts to pay more attention to populated area and tends to neglect the rare-event probability zone.

The proposed algorithm enables: (1) to substantially circumvent the abnormal ringing errors by eliminating the need of the inverse operation and (2) to guarantee the good enough deconvolution precision even if the shape of the RTN distribution is complex, comprising the complex gamma mixtures with the multiple convex and concave folding points.

To the best of our knowledge, there have been quite few papers to present the deconvolution algorithm for the SRAM-designs featuring an iterative partitioned forward-problem based deconvolution (PFDCV) process while comparing with the results based on the Richardson-Lucy algorithm.

8.3.1 Partitioned forward-problem based deconvolution

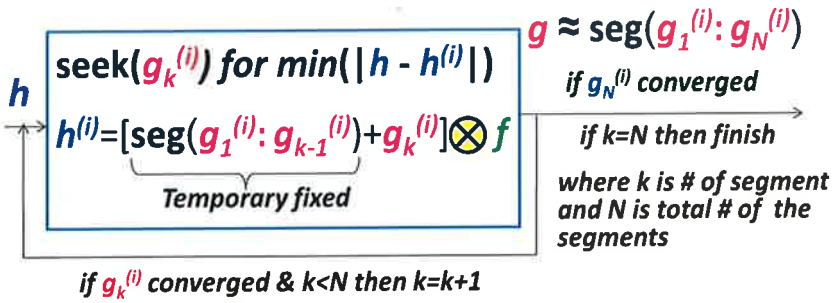


Figure 8.16: Proposed deconvolution algorithm featuring an iterative partitioned forward-problem based deconvolution (PFDCV) process.

Algorithm of the iterative partitioned forward-problem based procedure is described below from step (**i**) to (**iii**).

(1) The distribution of $\mathbf{g}^{(i)}$ is approximated by gamma distribution with three parameters of α (shape), β (inverse scale) and κ(peak value).

(2) Solving the optimization problem that seeks $\mathbf{g}^{(i)}$ for minimizing ($|\mathbf{h} - \mathbf{h}^{(i)}|$), where $\mathbf{h}^{(i)}$ is the convolution of RTN $\mathbf{g}^{(i)}$ with RDF **f**, then use "**fminsearch**" in MATLAB® to seek the best combination of (α, β, κ) for the approximation of $\mathbf{g}^{(i)}$ that minimizes the unconstrained multivariable function, which allows a derivative-free method.

(3) The $\mathbf{h}^{(i)}$ is defined as the convolution of the summation of the line-segment of ($\mathbf{g}_1^{(i)}$:$\mathbf{g}_{k-1}^{(i)}$) and $\mathbf{g}_k^{(i)}$ with **f**, where k is # of partition and N is total # of the partitions. i.e.,

$$\mathbf{h}^{(i)} = [\text{seg}(\mathbf{g}_1^{(i)}: \mathbf{g}_{k-1}^{(i)}) + \mathbf{g}_k^{(i)}] \otimes \mathbf{f} \qquad \text{---------(8.3)}$$

This flow can be repeated until k=N, as shown in Figure 8.16. The process of seeking the best $\mathbf{g}^{(i)}$ follows the sequentially step by step manner, i.e., from k=1 to k=N. Once found the best $\mathbf{g}_k^{(i)}$ in each segment, its value is temporally fixed when seeking the next $\mathbf{g}_{k+1}^{(i)}$ so that each optimization step cannot be interfered with by the other higher populated zone. This allows seeking the best $\mathbf{g}_N^{(i)}$ in the attention zone (k=N).

8.3.2 Concept of the proposed PFDCV method

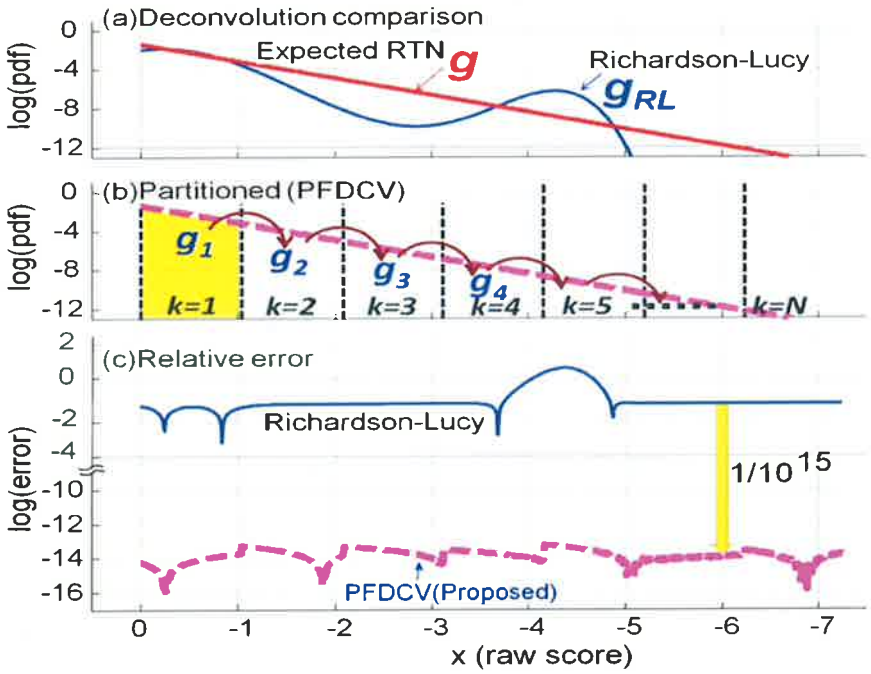


Figure 8.17: (a) RTN2 deconvolution comparison between the R-L and the expected one, (b) proposed algorithm (PFDCV) and (c) relative error comparison.

The concept of the proposed PFDCV method is illustrated in Figure 8.17(b). By avoiding the derivative operation, the behavior of the proposed RTN deconvolution process becomes

smoothed and stable. Figure 8.17(c) shows the comparisons of the relative deconvolution errors between the Richardson-Lucy and the proposed PFDCV. It is demonstrated that the proposed method can reduce the relative deconvolution error by 15-orders of magnitude compared with the Richardson-Lucy.

8.4 Discussion on Accuracy of Statistical Approximation Model for RTN Distribution

To illustrate the effects of the proposed scheme on the error reduction and convergence properties for the RTN1, RTN2, and RTN3, comparison results between the Richardson-Lucy and the proposed PFDCV are shown in Figure 8.18 and Figure 8.19, respectively.

It is found that the proposed PFDCV can reduce the deconvolution error for RTN1, RTN2, and RTN3 by 10^{11}, 10^{24}, 10^{14}-fold than that for Richardson-Lucy, as shown in Figures 8.18(a), 8.18(b), and 8.18(c), respectively.

The convergence properties of the iterative deconvolution process are compared between the Richardson-Lucy and the proposed PFDCV, as shown in Figure 8.19. This is the best advantage of the PFDCV over the Richardson-Lucy.

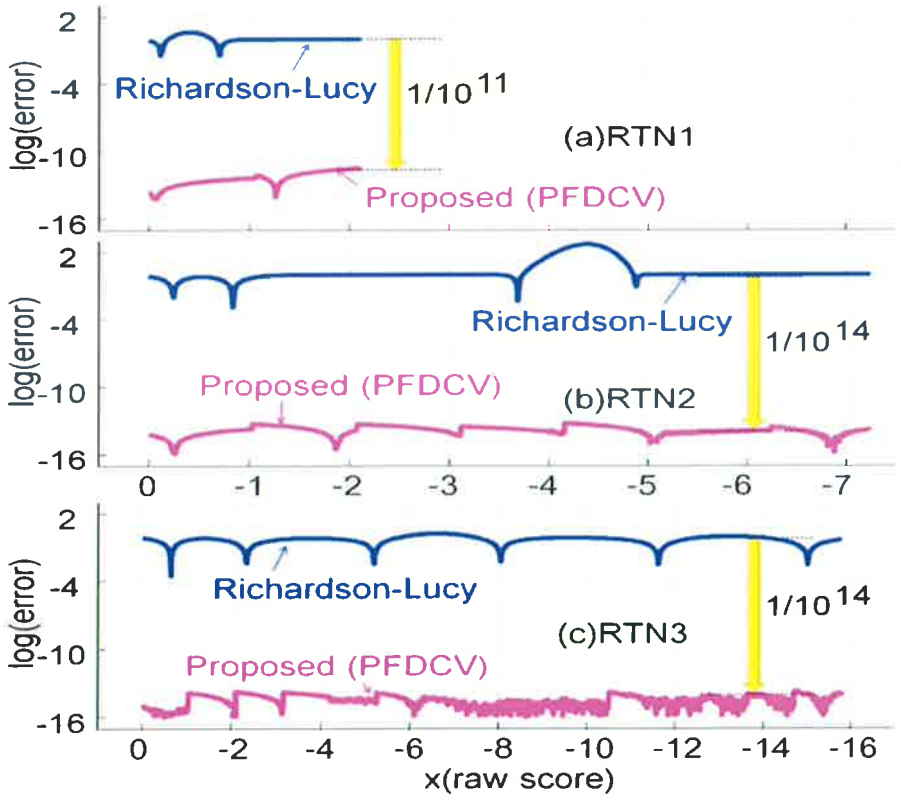


Figure 8.18: (a) RTN2 deconvolution comparison between the R-L and the expected one, (b) proposed algorithm (PFDCV) and (c) relative error comparison.

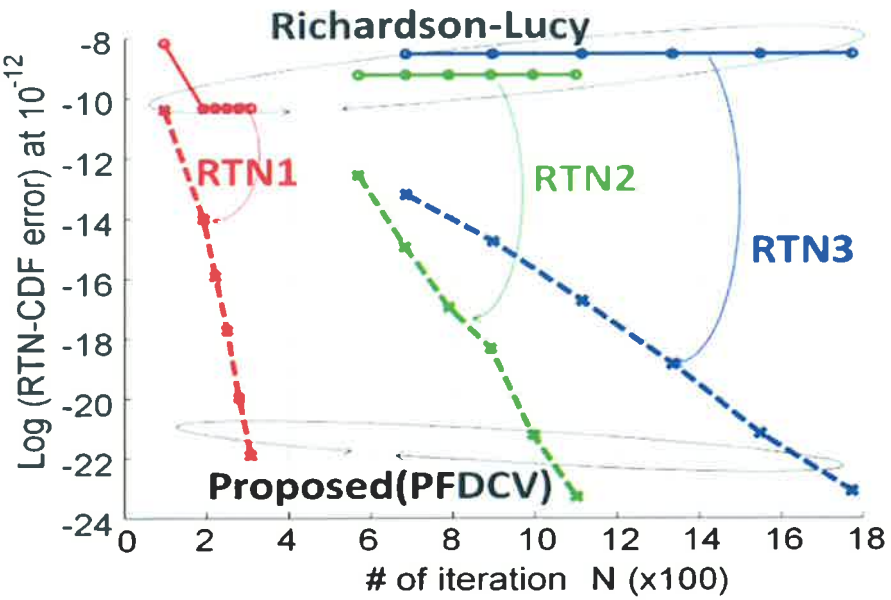


Figure 8.19: Comparisons of the convergence properties between the R-L and the proposed PFDCV.

8.5 Conclusions

In this chapter, we proposed the partitioned forward problem based deconvolution technique (PFDCV) enabling to successfully circumvent the issue of the ringing error confronting the Richardson-Lucy (R-L) process. The effectiveness of the PFDCV algorithm has been demonstrated for the first time with applying it to a real analysis for the effects of the RTN and the RDF on the overall SRAM margin variations.

The proposed PFDCV technique can reduce its relative RTN deconvolution errors by 10^{14}-fold compared with the cases of the Richardson-Lucy.

8.6 References

[1]. K. Takeuchi, T. Nagumo and T. Hase "Comprehensive SRAM Design Methodology for RTN Reliability", Symposium on VLSI Technology, pp. 130-131, June 2011.

[2]. K. Takeuchi, T. Nagumo, K. Takeda, S. Asayama, S. Yokogawa, K. Imai, Y. Hayashi, "Direct Observation of RTN-induced SRAM Failure by Accelerated Testing and Its Application to Product Reliability Assessment " , in Digest of IEEE Symposium on VLSI Technology 2010, pp. 189-190, 2010.

[3]. X. Wang, A.R. Broun, B. Cheng, A. Asenov, "RTS amplitude distribution in 20nm SOI FinFETs subject to Statistical Variability", Simulation of Semiconductor Processes and Devices SISPAD 2012, pp.296-299, 2012.

[4]. 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.

[5]. Q. Shi, O. Kanoun, "Application of Iterative Deconvolution for Wire Fault Location via Reflectometry", Digest of IEEE International Symposium on Instrumentation & Measurement, Sensor Network and Automation (IMSNA), pp. 102-106, 2012.

[6]. W. Somha, H.Yamauchi, "Convolution/ Deconvolution SRAM Analyses for Complex Gamma Mixtures RTN Distributions", Digest of ICICDT-2013, pp. 33-36, 2013.

[7]. W. Somha, H.Yamauchi, "A V-shaped Deconvolution Error Suppression Technique for SRAM Analyses of Long-Tail Gamma Mixtures Random Telegraph Noise Distribution Effects", Digest of ITC-CSCC-2013, pp. 297-299, 2013.

[8]. W. Somha, H.Yamauchi "Iterative and Adaptively Segmented Forward Problem Based Non-Blind Deconvolution Technique for Analyzing SRAM Margin Variation Effects", Digest of ISOCC-2013, pp. T.B.D. (accepted to appear), 2013.

[9]. Fish D.A. A.M. Brinicombe, E.R. Pike, "Blind deconvolution by means of the Richardson–Lucy algorithm", Journal of the Optical Society of America A12 (1), pp.58–65, 1995.

[10]. White R. L. "Image Restoration Using the Damped Richardson-Lucy Method" The Restoration of HST Images and Spectra II Space Telescope Science Institute, pp.104-110, 1994.

[11]. F. Dell Acqua1, "A Modified Damped Richardson-Lucy Algorithm to Improve the Estimation of Fiber Orientations in Spherical Deconvolution", Proc. Intl. Soc. Mag. Reson. Med. 16, pp. 1860, 2008.

Chapter 9

Comparative Study and Conclusions

To summarize this thesis, a brief summary for main achievements for the proposed EM algorithm and deconvolution algorithm is given followed by describing the prospective challenges and future work in this chapter. In order to clarify the effectiveness of the proposed algorithm, the results of comparative study between the proposed one and MATLAB built-in functions of Gaussian Mixture EM function and the various deconvolution functions are summarized in this chapter.

9.1 EM Algorithm for Approximation with Gaussian Mixtures

An effective simple algorithm for approximating long-tail distributions representing the future RTN variations related margin variables has been proposed. It has been shown that the proposed scheme can reduce not only the approximation error by 2-3 orders of magnitude but also required EM-step iterations.

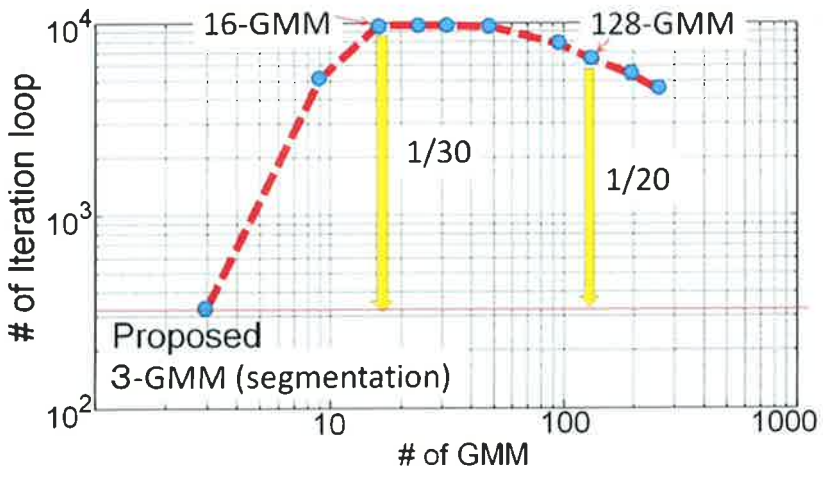


Figure 9.1: The number of iterations of EM steps is compared between 3-GMM segmentation schemes and conventional.

Since the reduction of the required iterations for EM step convergence is also crucial in practical application, the relationship between the number of GMM and the number of

iterations are compared, as shown in Figure 9.1 It is worth noting that the proposed 3-GMM segmentation scheme can reduce not only the error but also the EM-step iterations. The 3-GMM segmentation scheme can reduce by to 1/30 and 1/20 of that for 16 and 128-GMM without segmentation scheme.

In this work, we pointed out that the ordinary screening-test design (STD) will be no longer available when once the RTN could not be ignored any more, resulting from an intolerably increased effect on the shift-amount of the overall voltage-margin variations (OVMV). Since the major RTN effect on the OVMV modulation comes up after the screening test, the actual precision of the STD must rely entirely upon the estimation accuracy of the RTN effect.

When discussing the screening conditions for a volume production, the expected yield-loss should be assumed. This work assumed the target such that fail-probability is 10^{-12} to realize 99.9% yield of 1Gbit memory chip, which is a quite realistic target. Thus, attention point of x is around x=-16, where fail-probability is around 10^{-12}. In that sense, it can be said that our proposed 3-GMM segmentation method can provide the best approximation compared with others.

9.2 Comparative Study of Deconvolution of RTN and RDF

9.2.1 Comparisons of Algorithm and Its Errors

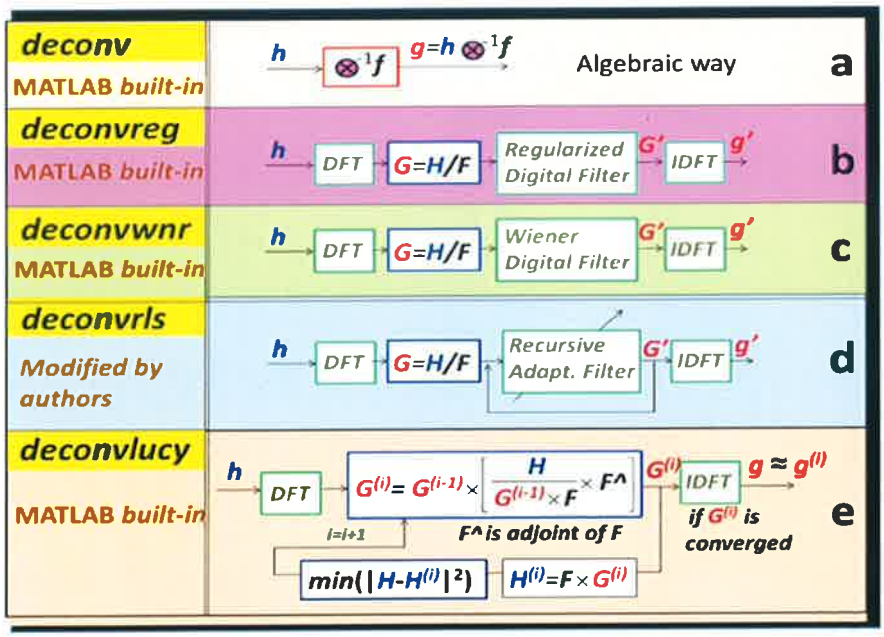


Figure 9.2: Various MATLAB® built-in deconvolution functions: (a) **deconv** (b) **deconvreg** (c) **deconvwnr** (e) **deconvlucy** and derivative functions made by authors (d) **deconvrls.**

MATLAB® (matrix laboratory) is a numerical computing environment and widely used in academic and research institutions as well as industrial enterprises. It has various built-in functions for the deconvolution, as shown in Figure 9.2.

"**deconv**" (see Figure 9.2(a)) is using long polynomial division to solve inverse problem. However, it has been reported that this induces the V-shaped ringing errors due to "division by zero" [8-11]. "**deconvreg**" and "**deconvwnr**" are using the regularized filter algorithm and Wiener filter algorithm, respectively (see Figure 9.2(b) and 9.2(c)). Both of the algorithms rely on digital filters based on a constrained optimization in the sense of least square error between the estimated **g'(x)** and the true one **g(x)** under requirement of preserving its smoothness. "**deconvrls**" (see Figure 9.2(d)) is not built-in MATLAB® but uniquely made by the authors with introducing the recursive adaptive filtering to reduce the noise. "**deconvlucy**" (see Figure 9.2(e)) uses Richardson-Lucy (R-L) algorithm [7-12] that is one of the most widely used deconvolution algorithms in the area of image processing although it has some shortcomings such as noise amplification [7-12]. This algorithm is based on maximizing the likelihood of the resulting **g(x)**, as shown in Figure 9.2(e).

The MATLAB® built-in deconvolution-functions are demonstrated for analyzing RTN effects on SRAM VCCmin-shift assuming the shape of distribution="**RTNcmb**".

Figures 9.3(a)-9.3(b) show the comparisons of the results of the deconvolution results **g'(x)=h(x)⊗$^{-1}$f(x)** among the cases of using 6-different algorithms including the proposed **"segmentation"** based deconvolution algorithm.

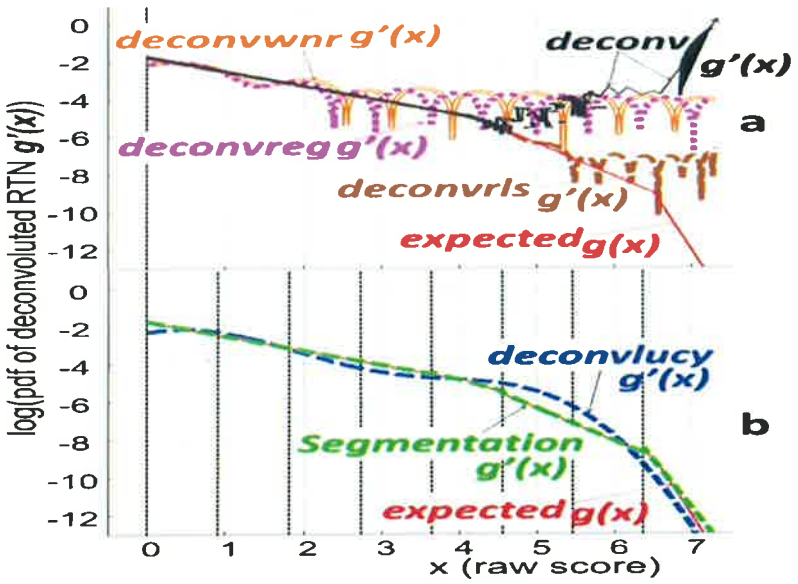


Figure 9.3: Comparisons of deconvolution for the different algorithms: "**deconv**", "**deconvreg**", "**deconvwnr**", "**deconvrls**", "**deconvlucy**" and proposed "**segmentation**".

"x(raw score)" is used as a title for the x-axis in Figure 9.3, and represents the scale of the distributions of the random variation, such as σVCCmin. The x-axis was normalized to address the different VCCmin distributions with Gaussian and Gamma distributions in the same figure. Y-axis is for log-scaled probability density function (pdf) of each deconvolution result.

It is found that both of deconvoluted **g'(x)** for "**deconvreg**" and "**deconvwnr**" exhibit similar ringing behavior in the x-range of x>2. Since that the range of ringing amplitude includes a negative value, some data are not plotted in log-scaled Y-axis. To address this issue, "**deconvrls**" uses a recursive adaptive filtering to reduce the noise, as shown in Figure 9.3(a). Thanks to the recursively adaptive update of the filtering characteristics, the starting point of ringing is significantly extended compared with "**deconvreg**" and "**deconvwnr**". However, it is not sufficient in the x-range of x>5, as can be seen in Figure 9.3(a).

Unlike the others, the Richardson-Lucy (R-L) algorithm based "**deconvlucy**" uses a forward problem based process to eliminate any division by zero issues. Instead, this seeks **g'(x)** based on the maximum likelihood iterations such that the difference between **h'(x)=f(x)⊗g'(x)** and **h(x)=f(x)⊗g(x)** becomes minimum. Thus, this can avoid a long polynomial division, resulting in a smaller level of the high-frequency noises compared with others, as can be seen in Figure 9.3(b). Figure 9.3(b) shows the remaining issue of low-frequency ringing error (noise-amplification) confronting the **"deconvlucy"** function. This is due to the maximum likelihood iterations with gradient method. As a result, the deconvoluted **g'(x)** is significantly deviated from the expected curve **g(x)** (see Figure 9.3(b)).

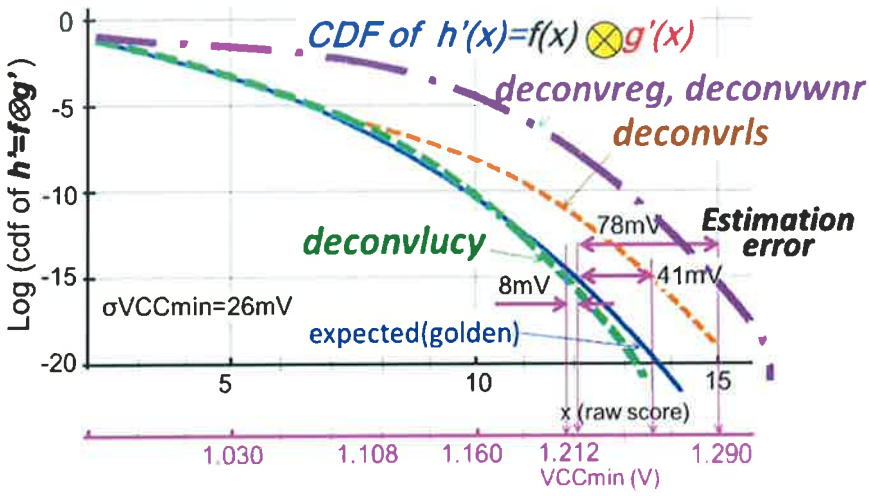


Figure 9.4: Comparisons of cdf error among "**deconv**", "**deconvreg**", "**deconvwnr**", "**deconvrls**", "**deconvlucy**" and golden.

The impact of deconvolution **g'(x)** errors on the prediction accuracy of VCCmin shift are compared. Figures 9.4 shows the comparisons of the cumulative density function (cdf) of the **h'(x)=f(x)⊗g'(x)**, which corresponds to the cdf distribution of VCCmin. The cdf lines given by "**deconvreg**" and "**deconvwnr**" are deviated from the golden-line (expected value) by a 78mV the golden-line (expected value). It is also found that "**deconvrls**" and "**deconvlucy**" induces the VCCmin prediction error by a 41mV and 8mV, respectively at the cdf=10^{-15}.

As explained so far, it is found that the MATLAB® built-in deconvolution-functions cannot be used for the VCCmin prediction of the nanometer scaled SRAMs because of an intolerable level of the errors not only for deconvoluted **g'(x)** but also **h'(x)=f(x)⊗g'(x)** in the tail region of its distributions. The main reasons for the errors are due to the noise amplification caused by long-division and maximum-likelihood (MLE) gradient iteration processes. Thus, to address this issue, a new deconvolution algorithm without using any long-division and MLE processes will be proposed in the following sections.

9.2.2 Comparisons of cdf errors Among Different Algorithms

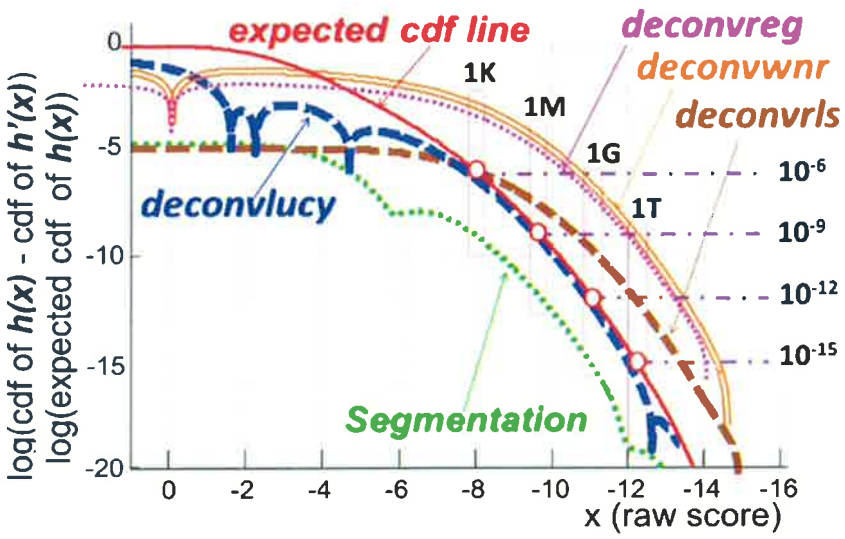


Figure 9.5: Comparisons of cdf-errors based on deconvoluted **g'(x)** RTNcmb among the cases of "**deconvreg**", "**deconvwnr**", "**deconvrls**", "**deconvlucy**" and proposed **"segmentation"**.

Figure 9.5 shows the comparisons of the cdf errors among the several cases with a different deconvolution algorithms, where, the cdf-error defines as |cdf(**h(x)**)-cdf(**h'(x)**)|. The box represents for the attention x-zone where its **"golden"** cdf line cross over the line of y=10^{-6}, 10^{-9}, 10^{-12}, 10^{-15}, i.e., corresponding to the probability point of "1-bit fail from 1000-pieces

of 1Kbit, 1Mbit, 1Gbit, and 1Tbit SRAMs". This zone is referred to as "attention x". Since the target of this work is to increase the accuracy of the FBC prediction, the error reduction of the cdf (x="attention x") is the key to the success of this work.

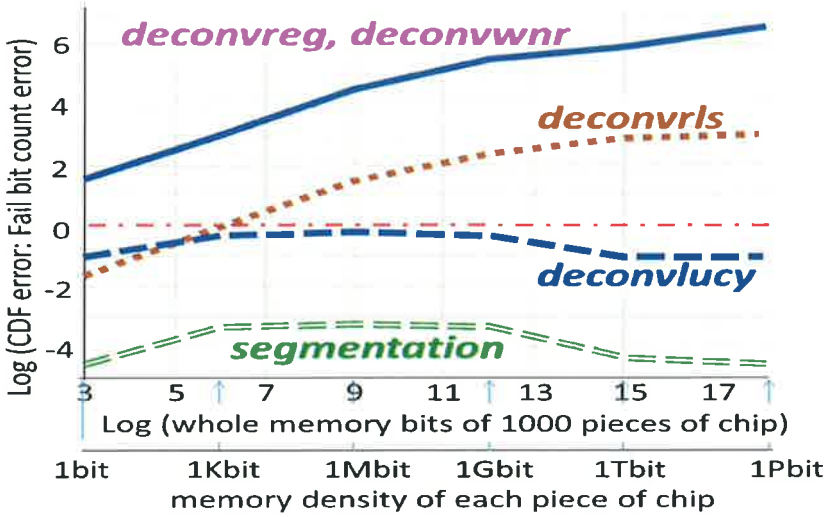


Figure 9.6: Comparisons of cdf-error compared with golden among the cases of "**deconvreg**", "**deconvwnr**", "**deconvrls**", "**deconvlucy**" and proposed "**segmentation**".

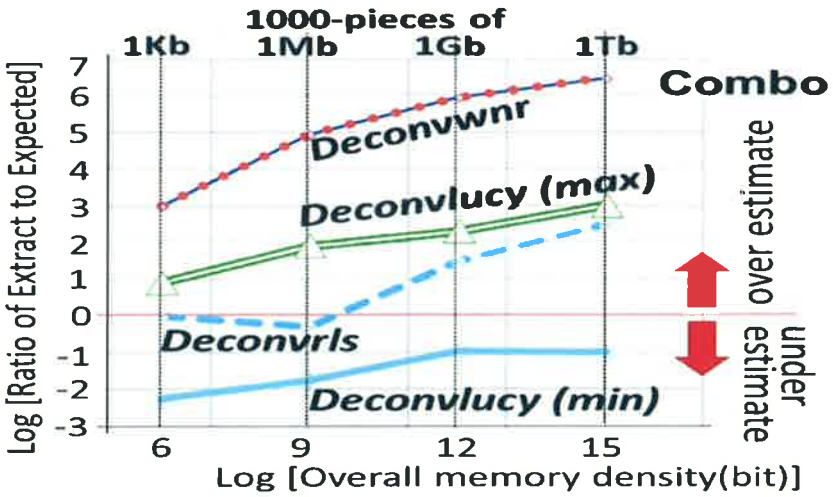


Figure 9.7: Comparisons of cdf-ratio among the cases of "**deconvreg**", "**deconvwnr**", "**deconvrls**", and "**deconvlucy**".

Figure 9.6 compares the cdf errors between the MATLAB® built in functions and the proposed segmentation one. It is found that the proposed algorithm can reduce the errors of the FBC prediction by about 10^3-folds compared with **"deconvlucy"** independent of the

memory bits ranging from 1000-pieces of 1Kbit to 1Tbit. On the other hand, the errors for **"deconvreg", "deconvwnr",** and **"deconvrls"** have a strong dependency of the memory density, as shown in Figure 9.6. Thus, the difference in the errors between the proposed **"segmentation"** and MATLAB built-in functions of **"deconvreg", "deconvwnr",** and **"deconvrls"** increases as the memory density is increased as shown in Figure 9.6.

Figure 9.7 shows the ratio of the cdf values, i.e., given by:

$$\text{Ratio}=\{\text{cdf(x) of h'(x)}\}/\{\text{cdf(x) of h(x)}\}. \qquad \text{--------(9.1)}$$

If the ratio value taken by log-scale is positive and negative values, it means "over-estimation" and "under-estimation", respectively. Log(ratio)=0 means error=0.

9.2.3 Effects of Adaptively Changing Segmentation Width

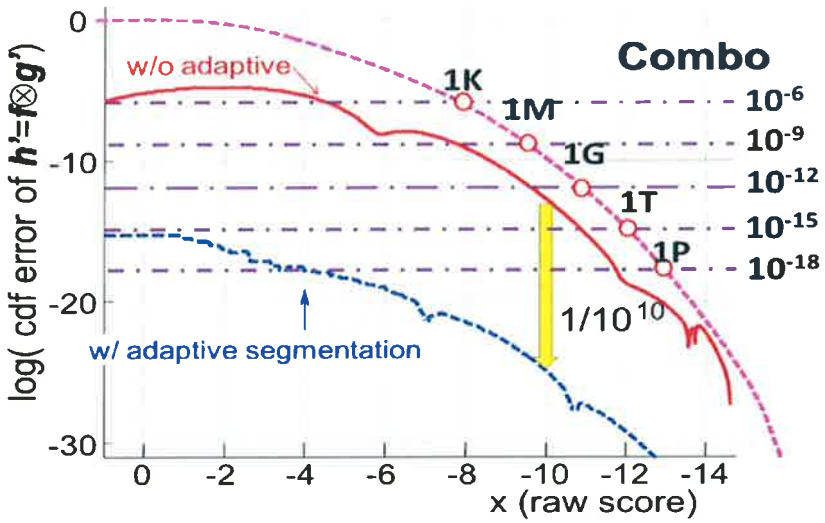


Figure 9.8: Comparisons of cdf-error compared with golden among the two cases for **"segmentation"** with and without adaptively changing segmentation width.

Figure 9.8 shows the comparisons of the cdf error of **h'(x)=f(x)⊗g'(x)** compared with golden cdf of **h(x)=f(x)⊗g(x)** among the two cases for **"segmentation"** with and without adaptively changing segmentation width. It is found that can the error be further reduced by 10^{10}-fold thanks to changing the segmentation width such that the folding points are set on the boundary line in the wide range of the memory densities from 1000-pieces of **1Kbit** to **1Pbit**, as shown in Figure 9.8.

The proposed **"segmentation"** with adaptively changing the segmentation width can further reduce the errors with increasing the iteration numbers. This is because allowance for increased numbers of iteration cycles can lead to reduce the segmentation width,

resulting in an increase in resolution of the deconvolution process. This is a clear advantage over other iteration algorithms of "**deconvrls**" and "**deconvlucy**", as can be seen in Figure 9.9.

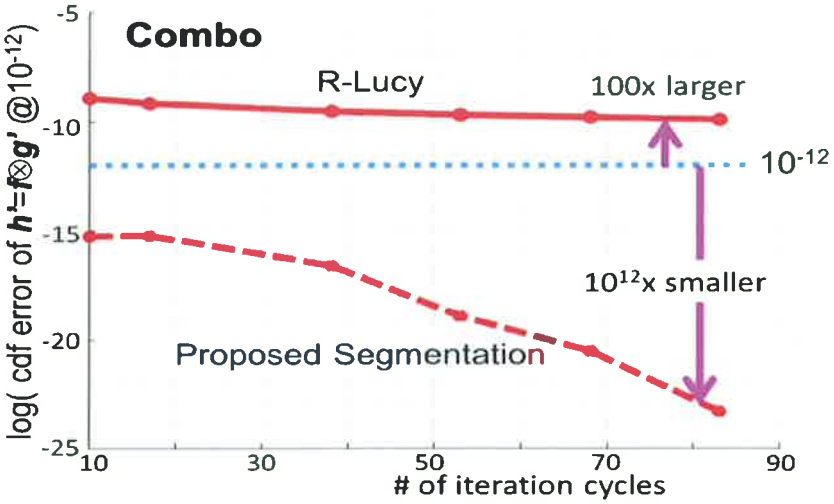


Figure 9.9: Comparison of the convergence characteristics of the error reduction when increasing the iteration cycles.

9.3 Prospective Challenges and Future Work

Table 9.1 Prospective challenges.

Item	Solution given by this study	Prospective challenges
EM algorithm for approximation and decomposition	Population-independent segmented approximation 1) Adaptive segmentation 2) Copy and paste fashion 3) Look up table based	Decomposing of the unknown kernel mixtures distribution with the EM-algorithm.
Deconvolution	Iterative forward based 1) Segmentation 2) +Least square regression	Iterative forward based 1) Required too much iteration 2) Ringing error in more complex mixtures cases still remained

In this study, EM algorithm for approximation with Gaussian mixtures and deconvolution techniques for the SRAM reliability screening test design have been proposed. Those can contribute not only for enabling to successfully circumvent the issue of the ringing error

confronting the MATLAB®built-in deconvolution-functions but also to improve the convergence capability of the error reduction process. However, some issues are not solved completely and unfortunately some challenges are remained, as summarized in Table 9.1.

1. This study prioritizes the increase of approximation accuracy with segmentation manner. However, decomposing of the complex distribution is also important role to make the deconvolution simpler. This requirement cannot be met well by this work.
2. This study prioritizes to avoid unstable convergence and unwanted ringing at the cost of iteration cycles. Ultimate requirements are to realize a fast and stable convergence while achieving deconvolution accuracy. Then modified Richardson-Lucy algorithm is still one of the attractive candidates.
3. Potential ideas for the modified Richardson-Lucy algorithm are divided into the two:(1) phase shift of the feedback gain distribution so that its phase can be aligned with the deconvolution object, and (2) segmentation for Richardson-Lucy algorithm. Another phase alignment techniques must be needed to eliminate any ringing for the complex deconvolutions for the multiple sloped complex segment mixtures. This is because the required phase shift amount must be dependent on the value of the slope (gradient).

Table 9.2 Future work.

Item	Prospective challenges	Future work
EM algorithm for approximation and decomposition	Decomposing of the unknown kernel mixtures distribution with the EM-algorithm.	New algorithm to decompose overall complex margin variations so that each decomposed element can be simply deconvoluted.
Deconvolution	Iterative forward based 1) Required too much iteration 2) Ringing error in more complex mixtures cases still remained	Iterative forward based 1) Iteration reducing algorithm 2) R-L with segmentation and phase shift algorithm

Acknowledgment

First and foremost I offer my sincerest gratitude to my advisor, Prof. Hiroyuki Yamauchi, who has supported me to complete thought-out my dissertation with his patience and knowledge while allowing me to work in my own way. I attribute the level of my doctor degree to his guidance, encouragement and effort. This dissertation would not have been completed without him.

The author would like to extend his special gratitude to Prof. Kazumasa Oida, Prof. Leonard Barolli and Prof. Kei Eguchi, the committee members for their valuable suggestions and very good comments about this thesis.

I am grateful to Assoc. Prof. Kitti Tirasesth, who gave me the chance to continue the study in doctoral level and prepared me for the challenges of studies in Japan, I also thank Fumie Yuki-sensei, for teaching Japanese language. A special thank goes to Mr. Mori Cho, Mr. Hiroshi Chikamatsu and all staff of International Affairs department of Fukuoka Institute of Technology (FIT) for their continuous support and help.

Yan Zhang and Ma Yuyu have helped me to correct Japanese documents and sometime we enjoyed talks and drinks for refreshments together in Japanese restaurant.

The Department of Information engineering FIT has provided the support and equipment which I used to complete my dissertation and all students at Yamauchi LAB have supported me.

Special gratefulness is extended to my parents and friends for their never-ending encouragement and patience.

Finally, I am grateful to King Mongkut's Institute of Technology Ladkrabang (KMITL) for supporting me by a scholarship throughout of my studies at FIT, Japan.

List of Papers

Journals Papers

1. **W. Somha**, H. Yamauchi, "An RTN Variation Tolerant SRAM Screening Test Design with Gaussian Mixtures Approximations of Long-Tail Distributions", Journal of Electronic Testing Theory and Applications., Vol.30, No.2, pp.171-181, March 2014, DOI 10.1007/s10836-014-5439-7, ISSN 0923-8174.

2. **W. Somha**, H. Yamauchi, "An Adaptively Segmented Forward Problem Based Non-Blind Deconvolution Technique for Analyzing SRAM Margin Variation Effects", Journal of Semiconductor Technology and Science, Vol.14, No.4, pp.365-375, August 2014, DOI 10.5573/JSTS.2014.14.4.365.

3. **W. Somha**, H. Yamauchi, M. Yuyu, "A Technique to Circumvent V-shaped Deconvolution Error for Time-dependent SRAM Margin Analyses", IEIE Transactions on Smart Processing & Computing, Vol. 2, No.4, pp.216-225, August 2013, DOI 10.5573/IEIESPC.2013.2.4.216, ISSN 2287-5255.

4. **W. Somha**, H. Yamauchi, "A Comparative Study on RTN Deconvolution of Richardson-Lucy and Proposed Partitioned Means for Analyzing SRAM Fail-Bit Prediction Accuracy", International Journal of Computer and Communication Engineering, Vol. 3, No. 3, pp.178-183, May 2014, DOI: 10.7763/IJCCE.2014.V3.315.

5. **W. Somha**, H. Yamauchi, "Adaptive Segmentation Gaussian Mixtures Models for Approximating to Drastically Scaled-Various Sloped Long-Tail RTN Distributions", International Journal of Future Computer and Communication, Vol. 2, No. 5, pp.407-412, October 2013, DOI: 10.7763/IJFCC.2013.V2.195.

6. **W. Somha**, H. Yamauchi, "A Discussion on RTN Variation Tolerant Guard Band Design Based on Approximation Models of Long-Tail Distributions for Nano-Scaled SRAM Screening Test ", International Journal of Computer and Electrical Engineering, Vol. 5, No. 4, pp.366-371, August 2013, DOI: 10.7763/IJCEE.2013. V5.733.

7. **W. Somha**, H. Yamauchi, "A Look up Table Based Adaptive Segmentation Gaussian Mixtures Model for Fitting Complex Long-Tail RTN Distributions ", International Journal of Materials, Mechanics and Manufacturing, Vol. 1, No. 3 pp.245-250, August 2013, DOI: 10.7763/IJMMM.2013.V1.52.

8. **W. Somha**, H. Yamauchi, "Fitting Mixtures of Gaussians to Heavy-Tail Distributions to Analyze Fail-Bit Probability of Nano-Scaled Static Random Access Memory", Advanced Materials Research Vol. 677 pp.317-325, February 2013, DOI:10.4028/www.scientific.net /AMR.677.317.

International Conference Papers

1. H. Yamauchi, **W. Somha**, " Errors in Solving Inverse Problem for Reversing RTN Effects on VCCmin Shift in SRAM Reliability Screening Test Designs", IEEE The 27th international IEEE SoC (System-on-Chip) Conference (SOCC 2014) , Las Vegas USA, September 2014, pp.318-323, DOI: 10.1109/SOCC.2014.(t.b.d).

2. H. Yamauchi, **W. Somha**, "A Technique to Solve Issue of Richardson-Lucy Deconvolution for Analyzing RTN Effects on SRAM Margin Variation ", 5th IEEE Latin American Symposium on Circuits and Systems, Santiago, Chile, February 2014, DOI 10.1109 LASCAS.2014.6820250.

3. **W. Somha**, H. Yamauchi, "More Stable and Faster Convergence Algorithm for Solving Deconvolution Problem of Peak-Shifted Gamma Mixtures RTN Distributions ", The 29th International Technical Conference on Circuit/Systems Computers and Communications (ITC-CSCC), Phuket, Thailand, pp.504-507, July 2014.

4. **W. Somha**, H. Yamauchi, "Iterative and Adaptively Segmented Forward Problem Based Non-Blind Deconvolution Technique for Analyzing SRAM Margin Variation Effects", International SoC Design Conference (ISOCC), Korea, pp.184-187, November 2013, DOI 10.1109 ISOCC.2013.6863967.

5. **W. Somha**, H. Yamauchi, M. Yuyu, "Blind Deconvolution Technique for Extracting Unknown Two Factors of RTN and Truncated RDF from Given Target for Overall SRAM Margin Variations", International SoC Design Conference (ISOCC), Korea, pp.188-191, November 2013,DOI 10.1109 ISOCC.2013.6863968.

6. **W. Somha**, H. Yamauchi, M. Yuyu, "A Technique to Circumvent Problematic Deconvolution Processes in Time-dependent SRAM Margin Analyses ", International Conference on Engineering, Applied Sciences, and Technology, Bangkok, Thailand, pp.11-16, August 2013.

7. **W. Somha**, H. Yamauchi, M. Yuyu, "A Comparative Review of Deconvolution Errors in SRAM Margin Analyses between Algebraic and Optimization Problem Based Approaches ", International Conference on Engineering, Applied Sciences, and Technology, Bangkok, Thailand, pp.60-65, August 2013.

8. **W. Somha**, H. Yamauchi, M. Yuyu, "A Discussion on SRAM Forward/Inverse Problem Analyses for RTN Long-Tail Distributions", IEEE Computer Society Annual Symposium on VLSI (ISVLSI 2013), Natal, Brazil, pp.58-63, August 2013, DOI 10.1109 ISVLSI.2013.6654623.

9. **W. Somha**, H. Yamauchi, "Convolution/Deconvolution SRAM Analyses for Complex Gamma Mixtures RTN Distributions", IEEE International Conference on IC Design and Technology, Pavia, Italy, pp.33-36, May 2013, DOI 10.1109ICICDT.2013.6563297.

10. **W. Somha**, H. Yamauchi, "A RTN Variation Tolerant Guard Band Design for A Deeper Nanometer Scaled SRAM Screening Test: Based on EM Gaussians Mixtures Approximations Model of Long-Tail Distributions", IEEE 14th Latin American Test Workshop, Cordoba, Argentina, May 2013, DOI 10.1109 LATW.2013.6562687.

2015.2-17.

山内寛行

2015.2.17