A Discussion on SRAM Forward/Inverse Problem Analyses for RTN Long-Tail Distributions

Worawit Somha,
Fukuoka Institute of Technology
Information Intelligent System3-30-1,
Wajiro-Higashi, Higashi-ku,
Fukuoka, Japan
e-mail: bd12002@bene.fit.ac.jp

Hiroyuki Yamauchi,
Fukuoka Institute of Technology
Information Intelligent System3-30-1,
Wajiro-Higashi, Higashi-ku,
Fukuoka, Japan
e-mail: yamauchi@fit.ac.jp

Ma YuYu,
Fukuoka Institute of Technology
Information Intelligent System3-30-1,
Wajiro-Higashi, Higashi-ku,
Fukuoka, Japan
e-mail: mfm12011@bene.fit.ac.jp

Abstract—This paper 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 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.

Keywords— Deconvolution, SRAM, Gamma Mixtures, Random telegraph noise, Long-tail distribution

I. 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 predicted 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 Figs. 1(a) and 1(b).

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 Fig.1(b)), 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 Figs. 1(c)-1(d) and Figs. 2(b)-2(d).

Fig. 1 (a) traditional and (b) new relationship between RDF, RTN, and its convolution. (c) and (d) show the deconvolution (inverse problem) required to detect the unknown factors.

Figure 2(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

MEXT/JSPS KAKENHI Grant Number of 23360424 and grant from Information Sceience Laboratory of Fukuoka Institute of Technology

U.S. Government work not protected by U.S. copyright
random telegraph noise (RTN). The examples for the inverse problems are shown in Figs. 2(b)-(d). Figure 2(b) recounts the following scenarios: target specs are predefined. The RTN distribution is also predefined or given, then, we 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 deconvolution analysis. Figure 2(c) assumes the different scenario: both of the target specs and the RTN distribution are predefined or given, then we want to know how much we need to shift the whole MV by using the MV assisted circuit schemes (ASSTS)[5]-[6]. Another scenario is shown in Fig. 2(d): the target specs and the truncated RDF distribution are predefined or given, then, we want to specify the device target of reducing the amplitude of RTN.

II. CONVOLUTION/DECONVOLUTION OF RTN AND RDF

It has been well shown in the references [4]-[6] that the distributions for the Vth modulation (ΔVth) amplitude due to RDF and RTN are obeyed to the Gaussian and Gamma distributions or more complex mixtures of them, respectively. The increasing paces of the ΔVth amplitude are differently dependent on the MOSFET size scaling parameter like below the expressions of (1) and (2).

\[
\Delta V_{th} \text{ (RDF)} \propto AV_t \text{ (RDF)} / \sqrt{LW} \quad (1)
\]
\[
\Delta V_{th} \text{ (RTN)} \propto AV_t \text{ (RTN)} / LW \quad (2)
\]

where AVt is Pelgrom coefficient and LW are MOSFET channel length and width, respectively. Assuming the LW is scaled down to 0.5 every process generation, the ΔVth increasing paces of RTN is a 1.4x faster than that of RDF. In addition, since the advanced CMOS device tends to change to much less-dopant body devices like FinFET, ultra-thin body SOI, and nanowire FET, the RDF increasing trends can be slow-down, resulting in a larger difference in the increasing pace than 1.4x. According to the references [1], there will come the time soon around a 15nm scaled CMOS era.

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-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.

| TABLE-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_n ⋆ g_1 |

(c) Proposed De-convolution

| Optimization problem 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_n ⋆ g_1) / f_1 |

A. Convolution computing

The convolution can be given by the equation shown in TABLE-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 \( f_1-f_n \) and \( g_1-g_n \) are given, the observation of \( h_1-h_n \) can be easily computed. This operation plays a role of low-pass filtering, resulting in \( h=f*g \) a smoothed curve.

**B. Issues facing deconvolution operations**

The deconvolution can be basically computed based on the equation shown in TABLE-1(b). If the observation of \( h_1-h_n \) and the parameter of \( f_1-f_n \) at the segmentation are given with the mathematical equation, the unknown parameter of \( g_1-g_n \) can be easily computed under the limited conditions.

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-1(b). This causes the issues facing the inverse operations: (1) division by zero and Gibbs phenomenon, resulting in a ringing curve. As can be seen in Fig. 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 Fig. 3(a) and 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 \( h: \sum(h_i: h_1, g_i) \)) as shown in TABLE-1(c). The unknown parameter \( g_1-g_n \) can be searched so that the difference between the given observation data of \( h_1-h_n \) and the objective function \( O(g) \) can be minimized.

**III. 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 (\( \Delta X \)) for RDF and RTN, the \( \Delta X \) 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 Figs. 4(a)-4(f). Each slope (O-P, P-Q, Q-R, R-S) is represented by the shape parameter \((\beta)\) 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.

**A. Convolution Results**

Fig. 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 \( f \).

Fig. 4 The relationships between the slope of RTN\((g)\) and the convolution\((h=f*g)\), (a)-(c): different single slopes (d)-(f): different slope mixtures

---

**TABLE-1(a)**

<table>
<thead>
<tr>
<th>RTN1 ((g))</th>
<th>RTN2 ((g))</th>
<th>RTN3 ((g))</th>
<th>RTN4 ((g))</th>
<th>RTN5 ((g))</th>
<th>RTN6 ((g))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Truncated RDF ((f))</td>
<td>Truncated RDF ((f))</td>
<td>Truncated RDF ((f))</td>
<td>Truncated RDF ((f))</td>
<td>Truncated RDF ((f))</td>
<td>Truncated RDF ((f))</td>
</tr>
</tbody>
</table>

**TABLE-1(b)**

<table>
<thead>
<tr>
<th>RTN1 ((g))</th>
<th>RTN2 ((g))</th>
<th>RTN3 ((g))</th>
<th>RTN4 ((g))</th>
<th>RTN5 ((g))</th>
<th>RTN6 ((g))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Truncated RDF ((f))</td>
<td>Truncated RDF ((f))</td>
<td>Truncated RDF ((f))</td>
<td>Truncated RDF ((f))</td>
<td>Truncated RDF ((f))</td>
<td>Truncated RDF ((f))</td>
</tr>
</tbody>
</table>

**TABLE-1(c)**

<table>
<thead>
<tr>
<th>RTN1 ((g))</th>
<th>RTN2 ((g))</th>
<th>RTN3 ((g))</th>
<th>RTN4 ((g))</th>
<th>RTN5 ((g))</th>
<th>RTN6 ((g))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Truncated RDF ((f))</td>
<td>Truncated RDF ((f))</td>
<td>Truncated RDF ((f))</td>
<td>Truncated RDF ((f))</td>
<td>Truncated RDF ((f))</td>
<td>Truncated RDF ((f))</td>
</tr>
</tbody>
</table>
As can be seen in the Figs. 4(a)-4(f), the convolution itself can be successfully done based on the equations of TABLE-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 \((g)\).

**B. New Deconvolution Results**

The deconvolution results for the RTN \((g)\) and RDF \((f)\) are shown in Figs. 5 and 6, respectively. The error of the deconvolution is defined as the relative error of the extrapolated value at each point, i.e., \(|g_i-g_{deconv}|/g_i\) and \(|f_i-f_{deconv}|/f_i\) 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 \(g_{deconv}\) is solved by the inverse problem (deconvolution of given \(h\) and \(f\)). As can be seen in Fig. 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 Fig. 6 are suppressed to \(<10^{-8}\). It is worth noting 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.

As explained in Fig. 2(b), the truncated point (TP) can be detected by the proposed methodology. The four types of the slope \((\beta)\) of RTN \((g)\) impacts on the convolution results \((f*g)\) are shown in Figs. 7(a)-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 Figs. 7(a) and 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 Figs. 7(c) and 7(d).

**IV. DETECTING THE TRUNCATED POINT**
The RTN-slope (β) dependencies of the shift amount (ΔX) of the convolution results (f*g) is shown in Fig. 8. It is worth noting 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 does work 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 Fig. 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.
V. ERRORS OF DECONVOLUTION

Figure 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 $10^{-13}$.

![Fig. 10 RTN-Slope ($\beta$) dependencies of the relative error of the RTN deconvolution.](image1)

VI. SUMMARY AND DISCUSSIONS

In this paper, we have discussed, for the first time, 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 have 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 demonstrated 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 have pointed out, for the first time, 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.

ACKNOWLEDGMENT

The authors are grateful to Yan Zhang for her helps.

REFERENCES


