A Comparative Study on RTN Deconvolution of Richardson-Lucy and Proposed Partitioned Means for Analyzing SRAM Fail-Bit Prediction Accuracy

Worawit Somha and Hiroyuki Yamauchi

Abstract—**This paper presents a comparative analysis of the Random Telegraph Noise (*RTN*) deconvolution accuracy between the Richardson-Lucy (R-L) algorithm and the proposed partitioned forward problem based deconvolution means (PFDCV). 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.**

Index Terms—**Random telegraph noise, fail-bit analysis, static random access memory, deconvolution, richardson-lucy deconvolution.**

I. INTRODUCTION

The approximation-error of the tails of random telegraph noise (*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 Fig. 1 and (2) the convolution result ($h=f \otimes g$) 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 ($h=f \otimes g$) between Fig. 2(a) and Fig. 2(b) for short and long *RTN*, respectively. Because the increasing paces of variation-amplitude σVth of the threshold voltage (*Vth*) are differently dependent on the MOSFET channel-size (*LW*) like the below expressions of (1) and (2), the σVth 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 Fig. 1.

$$\sigma Vth(RDF) \infty AVt(RDF)\big/\sqrt{LW} \qquad (1)$$

$$\sigma Vth(RTN) \infty AVt(RTN)\big/LW \qquad (2)$$

where *AVt* (*RDF*) and *AVt* (*RTN*) are Pelgrom coefficients for the *RDF* and the *RTN*, respectively.

According to the references [1]-[4], there will come the time soon around a 15nm-scaled CMOS era.
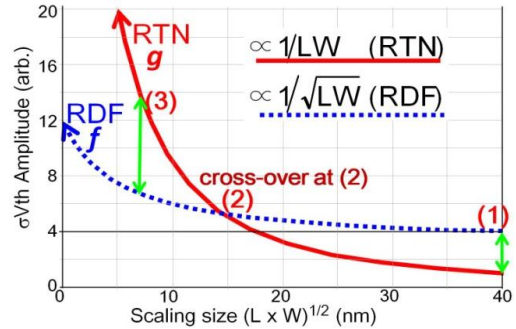
Fig. 1. Trend of variation amplitude of *RTN* and *RDF*. Variation amplitude of RTN becomes larger than that for *RDF* in 10nm era.
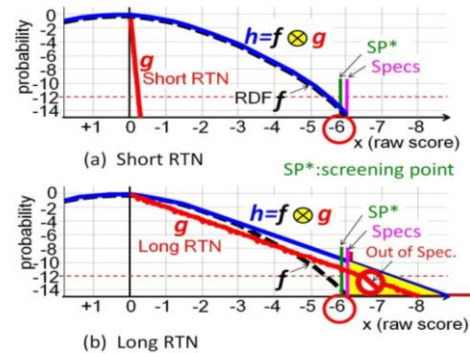

Fig. 2. Comparison of the convolution result $h=f \otimes g$ of the *RDF*(*f*) with *RTN*(*g*) between (a) short *RTN* and (b) long *RTN*. Convolution result h becomes governed by the *RTN*(*g*) when the tail length of g is larger than that for *f*.

The reliability 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 cannot be predicated any more by only ordinary *convolution* analyses [1]-[4].

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.

Fig. 3(a) and Fig. 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.

Fig. 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 f_{TP}. However, the *TP* of the f_{TP} and the random telegraph noise

(*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 SP_{prod}, as shown in Fig. 3(a). The *g* is calculated by the deconvolution of ($g=h\otimes^{-1}f_{TP}$). Unlike the convolution ($h=f_{TP}\otimes g$) (Fig. 3(b)), the deconvolution of the *g* is sort of ill-posed problem and troublesome operation [5]-[8].
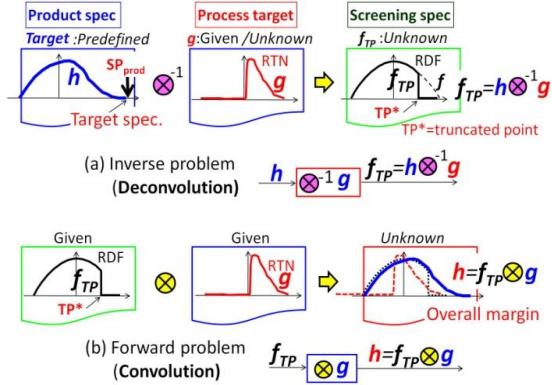

Fig. 3. (a) Inverse problem (deconvolution $\otimes^{-1}$) (b) forward problem (convolution $\otimes$).

Here is how the rest of this paper is organized. In Section II, we discuss the accuracy of the *RTN* deconvolution with Rechardson-Lucy algorithm. In Section III, we will propose the partitioned forward-problem based *deconvolution* (PFDCV) method. In Section IV, we rigorously prove that it is possible to reduce the deconvolution error with the proposed PFDCV method. Finally, we state our conclusion in Section VI.

II. DISCUSSIONS ON THE RICHARDSON-LUCY (R-L) DECONVOLUTION OF *RTN* ACCURACY

A. Richardson-Lucy Deconvolution Algorithm

Richardson-Lucy (R-L) algorithm [9] is one of the most widely used deconvolution algorithms (See Fig. 4) in the area of image processing although it has some shortcomings such as noise amplification [9]. As can be seen in Fig. 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.

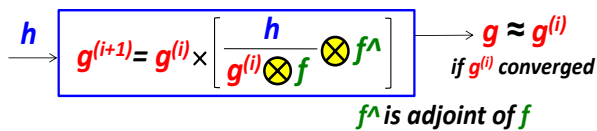As one of the tools for the deconvolution, the "*deconvlucy*" has been built in MATLAB®.

Fig. 4. Algorithm of Richardson-Lucy deconvolution of RTN(g).

However, to the best of our knowledge, there have been no examples of the R-L deconvolution being used for the SRAM margin analyses. It is for the first time to demonstrate the issues of the R-L deconvolution of the *RTN* distribution (*g*) with the SRAM MV (*h*). Fig. 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 Fig. 5(b))

As a result, the deconvoluted *RTN* distribution is significantly deviated from the expected curve (see Fig. 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}.
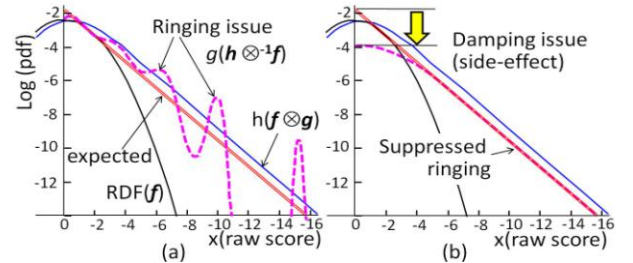

Fig. 5. Richardson-Lucy deconvolution of *RTN*(*g*) (a) ringing occurs (b) ringing is suppressed with damping factor=0.5 but the deconvolution pdf of *g* undershoots as a side-effect around *x*=0-4.

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.

B. Iteration Cycles Dependency of Deconvolution Errors

Fig. 6 shows the iteration cycle dependencies of the *RTN g* deconvolution with Richardson-Lucy (R-L) algorithm ($g_{RL}=h\otimes^{-1}f$). Where *h* is the convolution of the *RDF*(*f*) with the *RTN*(*g*), i.e., ($h=f\otimes g$). Relative R-L deconvolution error (g_{RL_ERROR}) is defined by the following expression of (3)

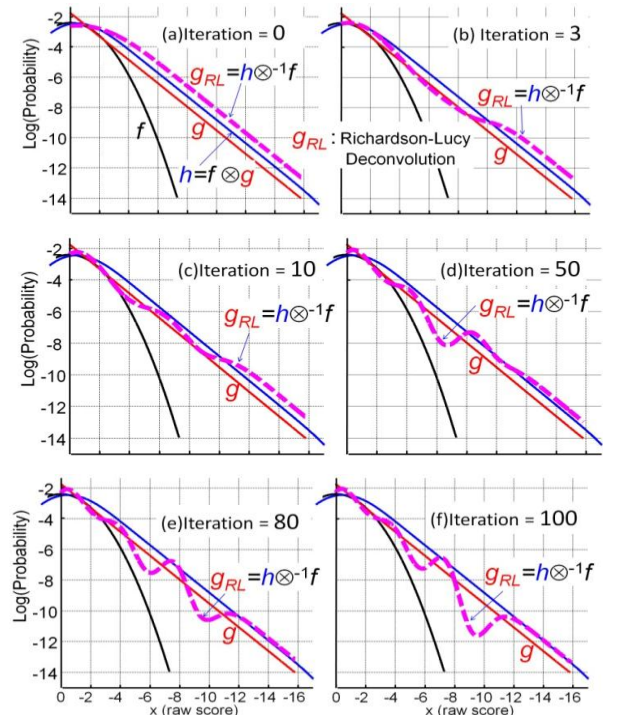$$g_{RL\_ERROR}=|\ g(x)-\ g_{RL}(x)|\ /\ g\ (x) \qquad (3)$$

Fig. 6. Iteration cycle dependencies of the deconvolution of *RTN g* ($g_{RL}=h\otimes^{-1}f$). where *N* is iteration numbers (a)*N*=0, (b)*N*=3, (c)*N*=10, (d)*N*=50, (e)*N*=80, and (f)*N*=100

As can be seen in Fig. 6(a)-Fig. 6(f), the *RTN* R-L deconvolution 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.

C. RTN Tail-length Dependency of Deconvolution Errors

As explained with Fig. 1, the amplitude of the *Vth* shift caused by the *RTN* is increased with the scaling trend. The *RTN*1, *RTN*2, and *RTN*3 (see Fig. 7(a), Fig. 7(b), and Fig. 7(c), respectively) correspond to the *RTN* distributions at the position of (1), (2), and (3) in Fig. 1, respectively. To make the *RTN*-length dependencies clearer, the R-L deconvolution of the *RTN* g_{RL} were compared among the *RTN*1, *RTN*2, and *RTN*3, while changing the iteration cycles N=10,100, and 1000, as shown in Fig. 8 and Fig. 9.
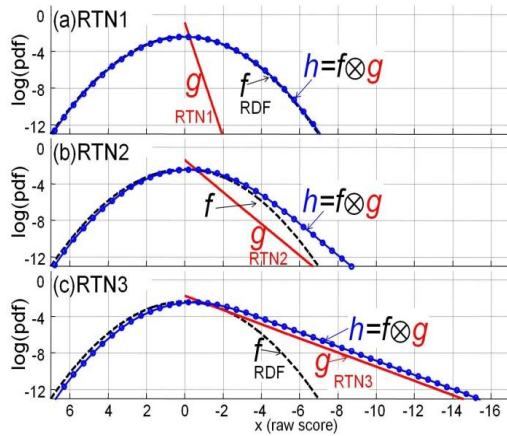

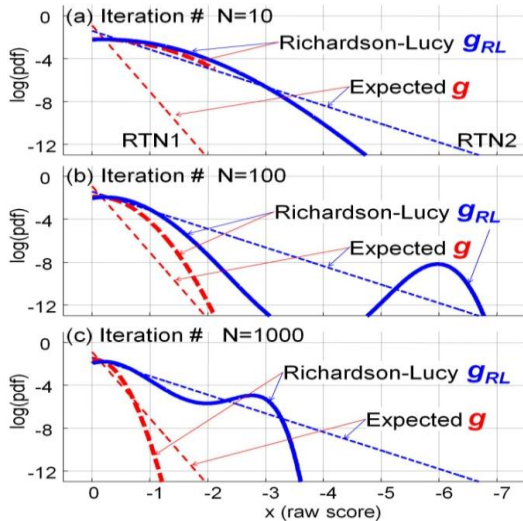Fig. 7. Relationship of the *RTN* tail-length among the *RTN*1,*RTN*2 and *RTN*3.

Fig. 8. Iteration cycle dependencies of the deconvolution of *RTN* g_{RL} ($g_{RL}=h\otimes^{-1}f$) for *RTN*1 and *RTN*2 for (a) N=10, (b) N=100, and (c) N=1000, respectively.

Fig. 8(a), Fig. 8(b), and Fig. 8(c) show the R-L deconvoltions of the *RTN*1 and the *RTN*2 at the iteration cycles N= 10, 100, and 1000, respectively.

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

for *RTN*1 in the region of *x*=0 to -1 is reduced while the error in the region of *x*=-1 to -2 is increased.

Fig. 9(a), Fig. 9(b), and Fig. 9(c) show the R-L deconvoltions of the *RTN*3 at the iteration cycles N= 10, 100, and 1000, respectively.

It is found that significant ringing happens in the R-L deconvolution of the *RTN*3 unlike the case for the *RTN*1 if the iteration number N is increased up to 1000. The relative error of the *RTN*3 deconvolution $g_{RL}(x)$ has a strong *x*-position dependency. As the N is increased, the relative error of $g_{RL}(x)$ for the *RTN*3 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.

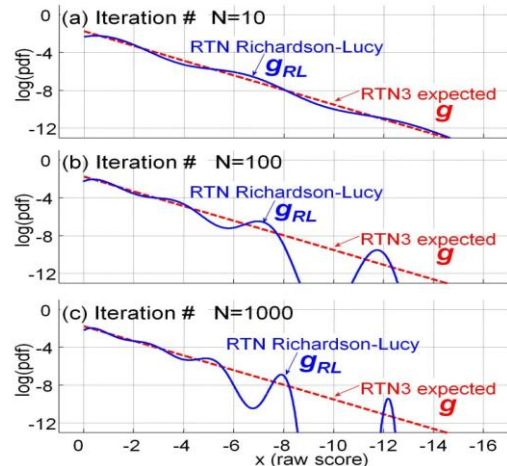Fig. 9. Iteration cycle dependencies of the deconvolution of *RTN* g_{RL} ($g_{RL}=h\otimes^{-1}f$) for *RTN*3 for (a) N=10, (b) N=100, and (c) N=1000, respectively.

D. 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 Fig. 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 Fig. 10(a). We refer this shape of the RTN distribution to "Combo". Fig. 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 Fig. 10(b).
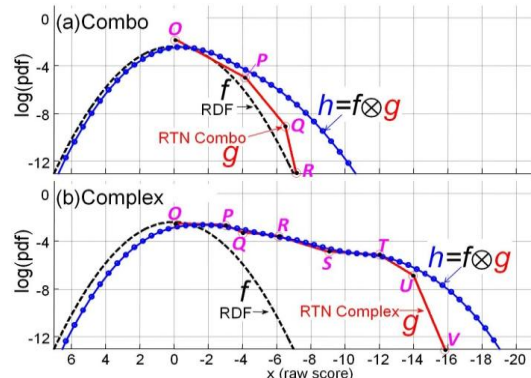

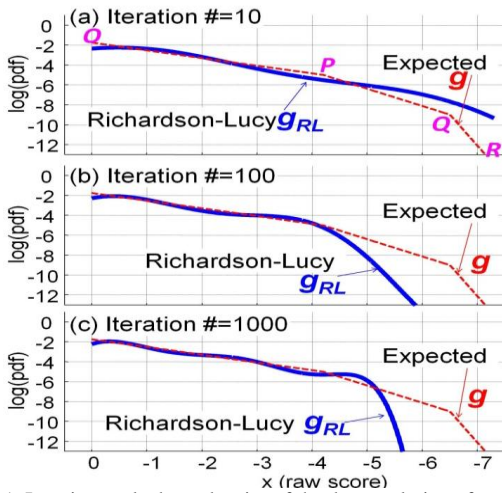Fig. 10. Relationship of the RTN tail-shape between "Combo" and "Complex".

Fig. 11. Iteration cycle dependencies of the deconvolution of *RTN* g_{RL} ($g_{RL}=h\otimes^{-1}f$) for "Combo" for (a) *N*=10, (b) *N*=100, and (c) *N*=1000, respectively.

Fig. 11(a), Fig. 11(b), and Fig. 11(c) show the R-L deconvoltions of the "Combo" at the iteration cycles *N* of 10, 100, and 1000, respectively.

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

Fig. 12(a), Fig. 12(b), and Fig. 12(c) show the R-L deconvoltions 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 $g_{RL}(x)$ has an x-position dependency. As the *N* is increased, the relative error of $g_{RL}(x)$ 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 *RTN*2, *RTN*3, and "Combo".
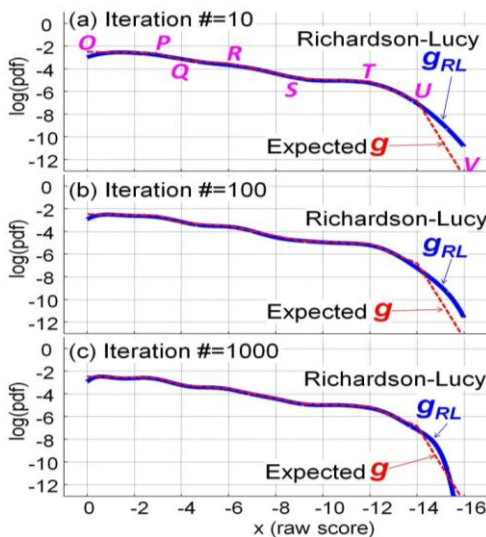

Fig. 12. Iteration cycle dependencies of the deconvolution of *RTN* g_{RL} ($g_{RL}=h\otimes^{-1}f$) for "Complex" for (a) *N*=10, (b) *N*=100, and (c) *N*=1000, respectively.

E. 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 *RTN*1, *RTN*2, *RTN*3, Combo, and Complex are compared, as shown in Fig. 13 and Fig. 14, respectively.
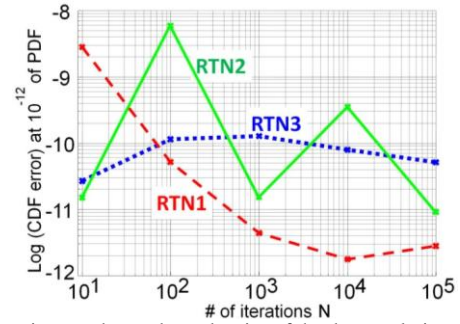

Fig. 13. Iteration number *N* dependencies of the deconvolution of *RTN* g_{RL} ($g_{RL}=h\otimes^{-1}f$) for *RTN*1, *RTN*2,
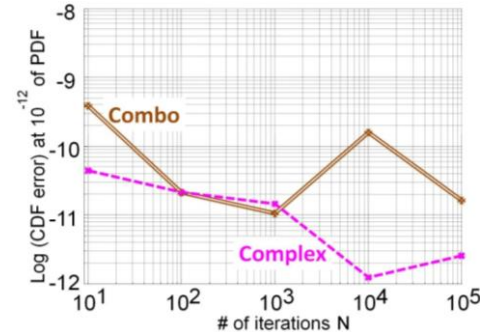

Fig. 14. Iteration number *N* dependencies of the deconvolution of RTN g_{RL} ($g_{RL}=h\otimes^{-1}f$) for "Combo" and "Complex".

It is found that the convergence properties for the *RTN*2 and the "Combo" are not secured. Error amplitude of the cumulative density fuction (*CDF*) are not converged but oscillated when the iteration number *N* is increased up to 10^5, as shown in Fig. 13 and Fig. 14. The convergence behavior is similar between the *RTN*2 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 Fig. 7(a) and Fig. 10(b).

Where, the error of the *CDF* is defined as the following expression (4)

$$CDF_{ERROR}=|CDF_{RL}(-Xp)-CDF(-Xp)|/CDF(-Xp) \qquad (4)$$

where CDF_{RL} is the *CDF* of deconvoluted *RTN* by the R-L algorithm. *Xp* is the point where *pdf*=10^{-12}

F. Damping Factor Dependency of Deconvolution Errors

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.
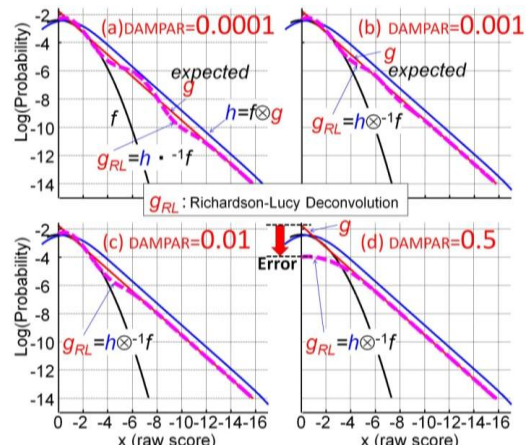

Fig. 15. Damping factor dependencies of the deconvolution of *RTN* g_{RL} ($g_{RL}=h\otimes^{-1}f$) for *RTN*3 *g*.

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 Fig. 15(d). The g_{RL} is deviated from the expected g of the *RTN*.

III. Proposed Partitioned Forward-Problem Based Deconvolution (PFDCV)

Thus, the proposed idea tries to *keep a sharp eye on* the rare event probability area by introducing the segmented optimization. This is where the proposed one is absolutely 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, this is the first time 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.

A. Partitioned Forward-Problem Based Deconvolution

Algorithm of the iterative partitioned forward-problem based procedure is described below from step (1) to (3).

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

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

3) The $h^{(i)}$ is defined as the convolution of the summation of the line-segment of ($g_1^{(i)}$:$g_{k-1}^{(i)}$) and $g_k^{(i)}$ with f, where k is # of partition and N is total # of the partitions. i.e.,

$$h^{(i)} =[\text{seg}(g_1^{(i)}: g_{k-1}^{(i)})+g_k^{(i)}] \otimes f \qquad (5)$$

This flow can be repeated until $k=N$, as shown in Fig. 16.

The process of seeking the best $g^{(i)}$ follows the sequentially step by step manner, i.e., from $k=1$ to $k=N$. Once found the best $g_k^{(i)}$ in each segment, its value is temporally fixed when seeking the next $g_{k+1}^{(i)}$ so that each optimization step cannot be interfered with by the other higher populated zone. This allows seeking the best $g_N^{(i)}$ in the attention zone ($k=N$).
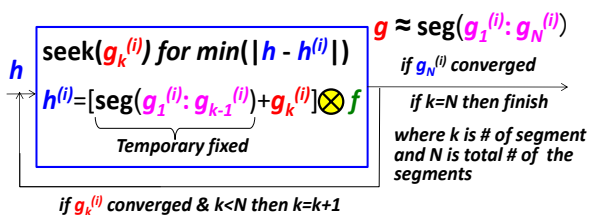

Fig. 16. Proposed deconvolution algorithm featuring an iterative partitioned forward-problem based deconvolution (PFDCV) process.

B. Concept of the Proposed PFDCV Method

The concept of the proposed PFDCV method is illustrated in Fig. 17(b). Thanks to avoiding the derivative operation, the behavior of the proposed *RTN* deconvolution process becomes smoothed and stable. Fig. 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.
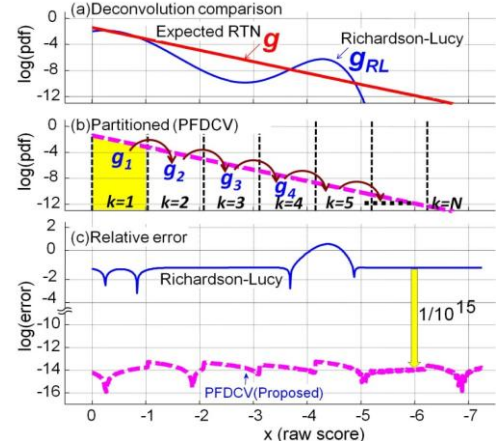

Fig. 17. (a) RTN2 deconvolution comparison between the R-L deconvolution and the expected one, (b) proposed deconvolution algorithm (PFDCV) and, (c) deconvolution relative error comparison between the R-L and the proposed PFDCV.

IV. Discusn on Accuracy of Statistical Approximation Model for RTN distribution

To illustrate the effects of the proposed scheme on 1) the error reduction and 2) convergence properties for the *RTN*1, *RTN*2, and *RTN*3, comparison results between the Richardson-Lucy and the proposed PFDCV are shown in Fig. 18 and Fig. 19, respectively.

It is found that the proposed PFDCV can reduce the deconvolution error for *RTN*1, *RTN*2, and *RTN*3 by 10^{11}, 10^{24}, 10^{14}-fold than that for Richardson-Lucy, as shown in Fig. 18(a), Fig. 18(b), and Fig. 18(c), respectively.

The convergence properties of the iterative deconvolution process are compared between the Richardson-Lucy and the proposed PFDCV, as shown in Fig. 19. This is the best advantage of the PFDCV over the Richardson-Lucy.
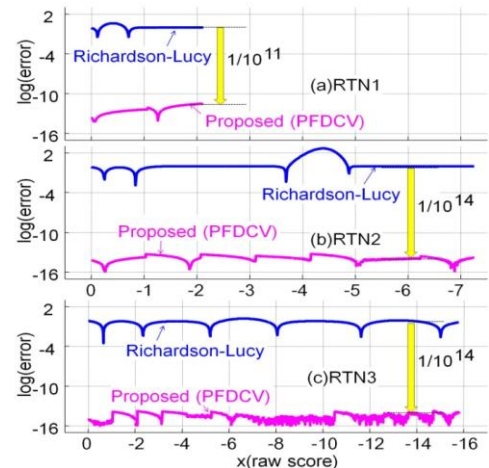

Fig. 18. Comparisons of the deconvolution error between the Richardson-Lucy and the proposed PFDCV.
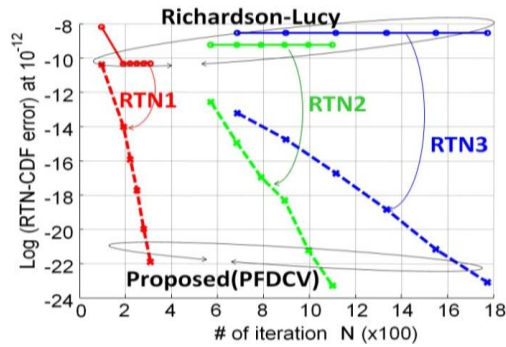
Fig. 19. Comparisons of the convergence properties between the Richardson-Lucy and the proposed PFDCV.

V. CONCLUSION

We have 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.

ACKNOWLEDGMENT

REFERENCES

[1] K. Takeuchi *et al*., "Comprehensive SRAM design methodology for RTN reliability," in *Proc. IEEE Symposium on VLSI Technology*, 2011, pp. 130-131.

[2] K. Takeuchi *et al*., "Direct observation of RTN-induced SRAM failure by accelerated testing and its application to product reliability assessment," in *Proc. IEEE Symposium on VLSI Tech*, 2010, pp. 189-190.

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

[4] X. Wang *et al*., "Simulation study of dominant statistical variability sources in 32-nm High-k/Metal Gate CMOS," *IEEE Electron Device Letters*, vol. 33, no. 5, pp. 643-645, 2012.

[5] Q. Shi *et al*., "Application of iterative deconvolution for wire fault location via reflectometry," in *Proc. IEEE Int. Symp. on Instrumentation & Measurement*, *Sensor Network and Automation*, 2012, pp. 102-106.

[6] W. Somha, H. Yamauchi, "Convolution/deconvolution SRAM analyses for complex gamma mixtures RTN distributions," in *Proc. Digest of ICICDT*, 2013, pp. 33-36.

[7] W. Somha and H. Yamauchi, "A V-shaped deconvolution error suppression technique for SRAM analyses of long-tail gamma mixtures random telegraph noise distribution effects," in *Proc. Digest of ITC-CSCC*, 2013, pp. 297-299.

[8] W. Somha and H. Yamauchi, "Iterative and adaptively segmented forward problem based non-blind deconvolution technique for analyzing SRAM margin variation effects," in *Proc. Digest of ISOCC*, 2013, pp. 184-187.

[9] D. A. Fish *et al*., "Blind deconvolution by means of the Richardson–Lucy algorithm," *Journal of the Optical Society of America A*, vol. 12, no. 1, pp. 58–65.

[10] R. L. White, "Image restoration using the damped richardson-lucy method," The Restoration of HST Images and Spectra II Space Telescope Science Institute, 1994, pp. 104-110.

[11] F. Dell'Acqua1, "A modified damped richardson-lucy algorithm to improve the estimation of fiber orientations in spherical deconvolution," in *Proc. Intl. Soc. Mag. Reson. Med.*, 2008, p. 16.

Worawit Somha received his master degree in electrical engineering from King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand. His master thesis was on "Vector quantizers for speech coding and there implementation on TMS-320C30". Since 1995 he has been a lecturer for bachelor degree students in subject of introduction to digital signal processing at KMITL as an assistance professor, and his research area is speech coding. Since 1997 he has worked in the position of consulting engineering. Since 2012 he won the scholarship from KMITL for the Dept. of Eng. student program and now being pursuing a PhD degree with a major in intelligence information system engineering at Fukuoka Institute of Technology. He is a member of the IEICE since 2012. He received the ISOCC2013 IEEK best paper award.

Hiroyuki Yamauchi received his Ph.D. degree in engineering from Kyushu University, Fukuoka, Japan, in 1997. His doctoral dissertation was on "Low power technologies for battery-operated semiconductor random access memories." In 1985 he joined the Semiconductor Research Center, Panasonic, Osaka, Japan. From 1985 to 1987 he worked on the research of the submicron MOS FET model-parameter extraction for circuit simulation and examined the sensitivity of the scaled sense amplifier for ultrahigh-density DRAM's which was presented at the 1989 Symposium on VLSI Circuits. From 1988 to 1994, he was engaged in the research and development of 16-Mb CMOS DRAM's including the battery-operated high-speed 16 Mbit CMOS DRAM and the ultra low-power, three times longer, self-refresh DRAM which were presented at the 1993 and 1995 ISSCC, respectively. He also presented the charge-recycling bus architecture and low-voltage operated high-speed VLSI's, including 0.5V/100MHz-operated SRAM and Gate-Over-Driving CMOS architecture, which were presented at the Symposium on VLSI Circuits in 1994 and 1996, respectively, and at the 1997 ISSCC as well. After being general manager for the development of various embedded memories, eSRAM, eDRAM, eFlash, eFeRAM, and eReRAM for system LSI in Panasonic, he moved to Fukuoka Institute of Technology and has been a professor since 2005. His current interests are focused on variation- tolerant memory circuit designs for the nanometer era. He holds 212 Patents including 87 U.S. Patents and has presented more than 70 journal papers and proceedings of international conferences including 10 for ISSCC and 11 for Symposium on VLSI Circuits. Dr. Yamauchi received the 1996 remarkable invention award from Science and Technology Agency of Japanese Government, and the highest ISOCC2008 best paper award and ISOCC2013 IEEK Best Paper Award.