Modeling Random Telegraph Noise as a Randomness Source and its Application in True Random Number Generation

Xiaoming Chen, *Member, IEEE*, Lin Wang, Boxun Li, *Student Member, IEEE*, Yu Wang, *Senior Member, IEEE*,
Xin Li, *Senior Member, IEEE*, Yongpan Liu, *Senior Member, IEEE*, and Huazhong Yang, *Senior Member, IEEE*

***Abstract*—The random telegraph noise (RTN) is becoming more serious in advanced technologies. Due to the unpredictability of the physical phenomenon, RTN is a good randomness source for true random number generators (TRNG). In this paper, we build fundamental randomness models for TRNGs based on single trap- and multiple traps-induced RTN. We theoretically derive the autocorrelation coefficient, bias, and bit rate for RTN-based TRNGs. Two representative RTN-based TRNG schemes are simulated to verify the proposed randomness models. An oscillator-based TRNG is also studied based on the theoretical randomness model of multiple traps-induced RTN. We also provide basic guidelines for designing RTN-based TRNGs.**

***Index Terms*—Random telegraph noise (RTN), randomness modeling, true random number generator (TRNG).**

I. Introduction

THE RANDOM telegraph noise (RTN) is a growing reliability issue in advanced integrated circuit (IC) technologies. RTN causes random fluctuations in electrical parameters such as the threshold voltage (V_{th}) and the source-drain current (I_{ds}). Recent studies have shown that at the 22 nm node, the RTN-induced V_{th} fluctuation can be larger than 70 mV, so RTN becomes a major noise source [1], [2]. RTN significantly affects the reliability of modern ICs. However, on the other hand, since the physical phenomenon of RTN is unpredictable, RTN potentially provides an excellent randomness source for creating true random number generators (TRNG). TRNG is an essential foundation in lots of cryptographic algorithms. Using pseudo random numbers will cause a big vulnerability because of the predictability. On the contrary, if the randomness source is unpredictable, like RTN, the generated random numbers will also be unpredictable. This is why TRNGs are essentially demanded for security.

A TRNG is typically composed of three major modules: 1) entropy source; 2) harvester; and 3) postprocessing. The entropy source provides raw random signals which are extracted from unpredictable physical phenomena. Raw signals are usually analog and biased. The harvester converts the raw signals to a digital bit stream. The post-processing is used to reduce bias to balance the probabilities of zeros and ones in the output.

A wide range of physical phenomena can be adopted to act as the entropy source, such as device noises, clock jitter, metastability, and chaos. It is claimed that some conventional entropy sources cannot offer high randomness, such as the clock jitter [3]. On the other hand, reliability mechanisms-based TRNGs are claimed to provide higher randomness [4], [5]. As a growing reliability mechanism, RTN offers much larger random fluctuations than well-known device noises, providing an excellent entropy source for TRNGs. In this paper, we will investigate the methodology of utilizing RTN as an entropy source in TRNGs. The purpose is to build fundamental randomness models for RTN and give basic guidelines to RTN-based TRNG design. To achieve this goal, we build systematic models to evaluate the randomness of RTN in theory. We also give basic methodologies to eliminate the autocorrelation and ensure high randomness for TRNGs based on both single trap- and multiple traps-induced RTN. Advantages of RTN-based TRNGs are emphasized by comparisons with conventional noise- and clock jitter-based TRNGs.

The rest of this paper is organized as follows. We review the related work and present our motivation in Section II. Statistical modeling and simulation methodology of RTN are introduced in Section III. In Sections IV and V, we present randomness models for single trap- and multiple traps-induced RTN, respectively. In Section VI, we study an oscillator-based

Manuscript received June 30, 2015; revised October 1, 2015; accepted December 10, 2015. Date of publication December 22, 2015; date of current version August 18, 2016. The work of Y. Wang was supported in part by the 973 Project under Grant 2013CB329000, in part by the National Natural Science Foundation of China under Grant 61373026 and Grant 61261160501, in part by the Tsinghua University Initiative Scientific Research Program, and in part by Huawei Technologies Company Ltd. The work of L. Wang was supported in part by the China Scholarship Council, and in part by the Capital University of Economics and Business under Grant 00591554410252. This paper was recommended by Associate Editor A. Raychowdhury.

X. Chen is with the Department of Electronic Engineering, Tsinghua National Laboratory for Information Science and Technology, Tsinghua University, Beijing 100084, China, and also with the Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213 USA (e-mail: chenxm1986@gmail.com).

L. Wang is with the School of Statistics, Capital University of Economics and Business, Beijing 100070, China (e-mail: wanglin@cueb.edu.cn).

B. Li, Y. Wang, Y. Liu, and H. Yang are with the Department of Electronic Engineering, Tsinghua National Laboratory for Information Science and Technology, Tsinghua University, Beijing 100084, China (e-mail: lbx13@mails.tsinghua.edu.cn; yu-wang@tsinghua.edu.cn; ypliu@tsinghua.edu.cn; yanghz@tsinghua.edu.cn).

X. Li is with the Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213 USA (e-mail: xinli@ece.cmu.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCAD.2015.2511074

TRNG using the proposed models. In Section VII, we compare RTN-based TRNGs with conventional noise- and clock jitter-based TRNGs. Finally, Section VIII concludes this paper.

II. Related Work and Motivation

In this section, we first briefly review some related work, and then present the motivation of this paper, followed by a summarized description of the proposed models.

A. Randomness Modeling of Noise

Kirton and Uren [6] gave a fundamental introduction on the physical origin and statistical characteristics of RTN based on the trap switching theory. Several studies built theoretical models for the jitter and phase noise in ring oscillators (ROs) by modeling the white noise and $(1/f)$ noise [7]–[9]. White noise was theoretically modeled to generate random numbers in [10].

For RTN, although many studies have derived statistical models based on measured data [11]–[15], they focus on modeling the physical phenomenon of RTN. Currently, there is no research that gives a systematic study on randomness modeling of RTN.

B. Existing TRNG Designs

According to the entropy source, there are several types of TRNGs. Entropy sources adopted by popular TRNGs include noises [16], [17], clock jitter in free-running ROs [18], [19], metastability [20], and chaos [21]. The breakdown mechanism of metal-oxide-semiconductor field-effect transistors (MOSFET) has also been studied to generate random numbers [4]. The only two RTN-based TRNGs are proposed in [5] and [22].

Almost all of these publications only provide implementations without any theory base on the randomness, naturally raising a question: is the randomness of such implementations really high enough? For example, the jitter-to-mean period ratio is at the magnitude of 10^{-4} in free-running ROs [18], such that the RO period must be very long to make a big jitter to create TRNGs. It is claimed that conventional noise- and metastability-based TRNGs cannot provide high randomness, due to the low magnitude of randomness or mismatch of devices [3]. In addition, whether a hard-to-describe chaotic system really behaves in a physically random fashion is unclear [3].

C. Motivation

As can be seen from the above, most of the existing TRNG designs lack for a theoretical derivation of the randomness. Utilizing RTN as a randomness source in TRNGs has two prominent advantages. As a growing reliability issue, RTN offers significantly large random fluctuations in advanced technologies, so that the fluctuations can be easily extracted and converted to random bits. In addition, more bits can be generated from each sampling due to the large fluctuation magnitude. We will further analyze the advantages of RTN-based TRNGs in Section VII. The physical phenomenon of RTN has been well modeled, but the randomness of RTN has never been systematically studied. This paper will build fundamental randomness models for RTN to provide a theoretical foundation for designing RTN-based TRNGs. We will focus on analyzing the autocorrelation coefficient, bias, and bit rate of RTN-based TRNGs. Among them, making a zero bias is the minimum requirement for random bits. However, only reducing the bias cannot guarantee the randomness. For example, a sequence $\{1, 0, 1, 0, 1, 0, 1, 0, \ldots\}$ has no bias but it is not random at all. Therefore, the autocorrelation which has a large impact on the randomness is also analyzed. A near-zero autocorrelation indicates that the next bit is difficult to predict when the previous bit is known. Bit rate is an important performance metric. Although there are other metrics or methods which can also be adopted to evaluate the randomness, such as the information entropy and the test suite provided by the National Institute of Standards and Technology (NIST) [23], they are high-level scores without tight connections with the physical characteristics of random bits. We choose autocorrelation, bias, and bit rate because they have clear physical meanings and crucial influences on the randomness and performance of TRNGs.

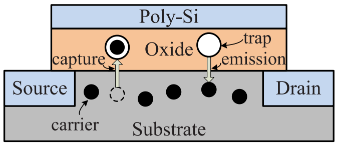


Fig. 1. Origin of RTN: the capture/emission process of traps.

D. Key Points of the Proposed Models

For TRNGs based on single trap-induced RTN, we will show how to select the sampling frequency to ensure a small autocorrelation. The maximum sampling frequency is constrained by the time constants of the trap. The bias can be eliminated by post-processing like the von Neumann corrector [24]. For TRNGs based on multiple traps-induced RTN, the sampling frequency can be close to the switching frequency of the fastest trap. We will show how to design a bit truncation scheme to eliminate the bias and the high autocorrelation.

III. Statistical Modeling and Simulation Methodology of RTN

In this section, we introduce existing statistical models and our simulation methodology of RTN.

A. Statistical Modeling of RTN

1) Physics of RTN: RTN can be explained by the random capture/emission process of charge carriers caused by oxide traps [25], as shown in Fig. 1. A trap in the oxide can occasionally capture a charge carrier from the channel, and the captured carrier can be emitted back to the channel after a period of time. The duration time of the captured and emitted states are denoted as τ_e (time before emission) and τ_c (time before capture), respectively, as marked in Fig. 2(a). In the time domain, V_{th} shows a binary fluctuation caused by a single trap. In the frequency domain, the power spectrum density (PSD) of the RTN-induced V_{th} fluctuation shows a

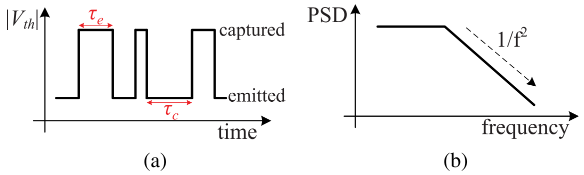


Fig. 2. RTN behavior in the (a) time and (b) frequency domains.

Lorentzian shaped spectrum with a slope of $(1/f^2)$ [6], as shown in Fig. 2(b).

2) Time Constants: The switching process of an individual trap obeys a Poisson process [6]. As a result, the duration time of the captured and emitted states follow exponential distributions with mean values $\bar{\tau}_c$ and $\bar{\tau}_e$, respectively [12]:

$$f(\tau_c) = \frac{1}{\bar{\tau}_c} e^{-\frac{\tau_c}{\bar{\tau}_c}} f(\tau_e) = \frac{1}{\bar{\tau}_e} e^{-\frac{\tau_e}{\bar{\tau}_e}} \tag{1}$$

where $\bar{\tau}_c$ and $\bar{\tau}_e$ are called the capture and emission time constants, respectively. They have a wide range from microsecond to millisecond when sampling a number of traps [11], [12]. Time constants of numerous traps can be approximated modeled by uniform distributions in the logarithmic scale [6], [12], [26]

$$\log_{10}(\bar{\tau}_c) \sim U(A_c, B_c), \log_{10}(\bar{\tau}_e) \sim U(A_e, B_e) \tag{2}$$

where $U(A, B)$ denotes the uniform distribution in the interval (A, B). Equation (2) is a statistical model for a number of traps. For each individual trap, its $\bar{\tau}_c$ and $\bar{\tau}_e$ are strongly correlated [12]. Although an analytical relation between $\bar{\tau}_c$ and $\bar{\tau}_e$ has been given in [6] and [11], it relies on some low-level parameters which are difficult to obtain and model. For convenience, the model can be simplified to

$$\frac{\bar{\tau}_e}{\bar{\tau}_c} = 10^m, m \sim U\left(\bar{m} - \frac{\sigma_m}{2}, \bar{m} + \frac{\sigma_m}{2}\right) \tag{3}$$

where $\bar{m}$ and σ_m are fitting parameters. $\bar{m}$ is linear to the bias voltage V_{gs} which indicates that $\bar{\tau}_c$ and $\bar{\tau}_e$ are approximately exponential to V_{gs} [11]. Using $\sigma_m = 2$ can generally fit the silicon data presented in [12]. The randomness of m denotes the statistical characteristics of trap positions and energies.

The above model is applicable in the case of a constant bias condition. Actually, $\bar{\tau}_c$ and $\bar{\tau}_e$ both depend on the bias condition. When a transistor undergoes periodically alternate two states (ST_1 and ST_2) with a fixed duty cycle, the periodic behavior of RTN can be described by an equivalent stationary RTN with two equivalent time constants [13]

$$\frac{1}{\bar{\tau}_c^{(\mathrm{equ})}} = \frac{\alpha}{\bar{\tau}_c^{(\mathrm{ST}_1)}} + \frac{1-\alpha}{\bar{\tau}_c^{(\mathrm{ST}_2)}}, \frac{1}{\bar{\tau}_e^{(\mathrm{equ})}} = \frac{\alpha}{\bar{\tau}_e^{(\mathrm{ST}_1)}} + \frac{1-\alpha}{\bar{\tau}_e^{(\mathrm{ST}_2)}} \tag{4}$$

where α is the duty cycle. $\tau_c^{(\mathrm{equ})}$ and $\tau_c^{(\mathrm{equ})}$ are fixed if α is fixed. Consequently, the model with fixed time constants can also be used in the case of a cyclostationary state.

3) Number of Traps: For numerous transistors, the number of detectable traps in each transistor statistically obeys a Poisson distribution [14]

$$P(N_t = k) = \frac{\langle N \rangle^k e^{-\langle N \rangle}}{k!} \tag{5}$$

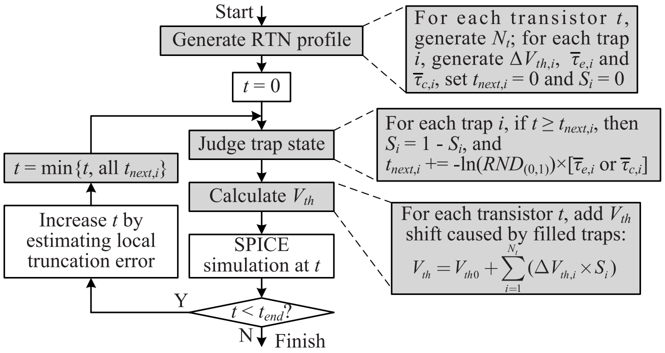

Start
Generate RTN profile
For each transistor t, generate N_t; for each trap i, generate $\Delta V_{th,i}$, $\overline{\tau}_{e,i}$ and $\overline{\tau}_{c,i}$, set $t_{next,i} = 0$ and $S_i = 0$
$t = 0$
Judge trap state
For each trap i, if $t \geq t_{next,i}$, then $S_i = 1 - S_i$, and $t_{next,i}$ += -ln($RND_{(0,1)}$)×[$\overline{\tau}_{e,i}$ or $\overline{\tau}_{c,i}$]
$t = \min\{t, \text{all } t_{next,i}\}$
Calculate V_{th}
For each transistor t, add V_{th} shift caused by filled traps:
$V_{th} = V_{th0} + \sum_{i=1}^{N_t} (\Delta V_{th,i} \times S_i)$
Increase t by estimating local truncation error
SPICE simulation at t
Y
$t < t_{end}$?
N
Finish

Fig. 3. SPICE-based simulation flow for RTN.

where N_t is the number of detectable traps in a transistor t, and $\langle N \rangle$ is the mean value of the Poisson distribution. pMOSFETs have more traps than nMOSFETs [14]. $\langle N \rangle$ increases with the shrinking of the feature size [27].

4) V_{th} Amplitude: The RTN-induced V_{th} fluctuation of numerous traps statistically obey an exponential distribution [15]

$$f(\Delta V_{\mathrm{th},i}) = \frac{1}{\langle \Delta V_{\mathrm{th}} \rangle} e^{-\frac{\Delta V_{\mathrm{th},i}}{\langle \Delta V_{\mathrm{th}} \rangle}} \tag{6}$$

where $\Delta V_{\mathrm{th}}, i$ is the V_{th} fluctuation caused by a trap i, and $\langle \Delta V_{\mathrm{th}} \rangle$ is the mean value of the exponential distribution. $\langle \Delta V_{\mathrm{th}} \rangle$ increases with the shrinking of the feature size [15]. In this paper, we set $\langle \Delta V_{\mathrm{th}} \rangle$ according to the 22 and 32 nm silicon data presented in [2] and [28]. The total V_{th} fluctuation of a transistor t is the superposition of the effects of all the individual traps in the transistor [29]

$$\Delta V_{\mathrm{th}} = \sum_{i=1}^{N_t} \left(\Delta V_{\mathrm{th},i} \times S_i\right) \tag{7}$$

where $S_i \in \{0, 1\}$ indicates the state of the ith trap in the transistor (0 for the emitted state and 1 for the captured state).

B. Simulation Methodology

Since currently there is no available circuit simulator which can natively support RTN, we build an in-house simulator for RTN estimation. Our simulator is a standard simulation program with IC emphasis (SPICE)-based tool [30] with the BSIM4 [31] model integrated. We follow the approach proposed in [29] to integrate the RTN models presented in Section III-A into the simulator. The simulation flow is shown in Fig. 3, where shaded blocks are RTN-related.

Before transient simulation, the RTN profile is given or randomly generated using the following three steps.

1) The number of traps in each transistor N_t is given, or generated by (5), where $\langle N \rangle$ is given.
2) The V_{th} fluctuation of each trap $\Delta V_{\mathrm{th},i}$ is given, or generated by (6), where $\langle \Delta V_{\mathrm{th}} \rangle$ is given.
3) The capture time constant of each trap $\bar{\tau}_{c,i}$ is given, or generated by (2), where the range is given. The emission time constant of each trap $\bar{\tau}_{e,i}$ is given, or generated by (3), where $\bar{m}$ is given.

Note that each trap has its own time constants and V_{th} amplitude [29]. As shown in Fig. 3, each trap i keeps a parameter

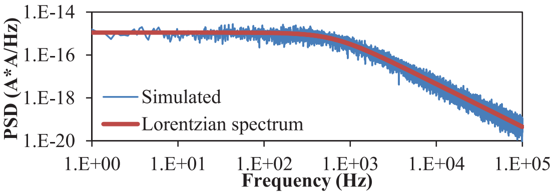


Fig. 4. Simulated and theoretical PSDs of I_{ds} fluctuation caused by single-trap induced RTN.

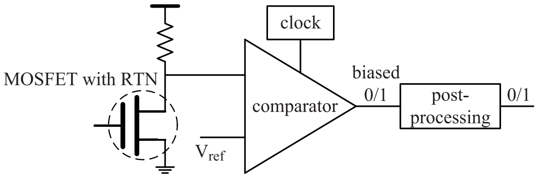


Fig. 5. Representative scheme for generating random numbers from single trap-induced RTN.

$t_{\text{next},i}$ which indicates when it will change its state S_i. During transient simulation, once the time point t reaches $t_{\text{next},i}$, trap i changes its state S_i, and then $t_{\text{next},i}$ is updated by the duration time of the next state, which obeys an exponential distribution. In simulation, the duration time of the next state is randomly generated by the following approach [29]:

$$\tau_c = -\ln(\text{RND}_{(0,1)}) \times \bar{\tau}_c,\ \tau_e = -\ln(\text{RND}_{(0,1)}) \times \bar{\tau}_e \qquad (8)$$

where $\text{RND}_{(0,1)}$ denotes a uniformly distributed random number in the interval $(0, 1)$. The RTN-induced ΔV_{th} of each transistor is calculated by (7) and added to the original V_{th} of each transistor, and then the BSIM4 model evaluation is performed based on the updated V_{th}. To avoid missing any state change during transient simulation, the time point t is always not larger than the minimum value of all the $t_{\text{next},i}$'s.

The accuracy of our simulator have been verified by comparing waveforms with commercial tools. Here, we verify the simulated PSD of single trap-induced RTN. In this test, we use the 22 nm high-performance predictive technology model (PTM) [32] in our netlists. A circuit with a single nMOSFET (the width is 50 nm) which has a single trap-induced RTN effect is simulated. The nMOSFET is stressed by a fixed bias condition of $V_{\text{gs}} = 0.6$ V and drives a load resistance of 10 kΩ. We use a fixed RTN profile in this test: $\bar{\tau}_c = \bar{\tau}_e = 0.5$ ms, $\Delta V_{\text{th}} = 20$ mV, and $N_t = 1$. Fig. 4 shows the simulated PSD of I_{ds} and the theoretical Lorentzian PSD which is given by [6]

$$S(f) = \frac{4(\Delta I_{ds})^2 \tau_0^2}{\bar{\tau}_c + \bar{\tau}_e} \cdot \frac{1}{1 + (2\pi f \tau_0)^2} \qquad (9)$$

where

$$\frac{1}{\tau_0} = \frac{1}{\bar{\tau}_c} + \frac{1}{\bar{\tau}_e}. \qquad (10)$$

$\Delta I_{ds} \approx 2.1\ \mu$A is obtained from the simulated I_{ds} waveform. Fig. 4 proves that the simulated PSD is well consistent with the theoretical Lorentzian spectrum.

IV. Modeling Single Trap-Induced RTN

In this section, we first derive a theoretical randomness model for single trap-induced RTN, and then analyze the performance of a representative TRNG scheme based on single trap-induced RTN.

Since the fluctuation caused by single trap-induced RTN has only two discrete values, the simplest way to generate random numbers is to convert the fluctuation into binary bits by a periodically sampled comparator, just like the approach proposed in [5]. Fig. 5 shows a representative scheme for this method. An amplifier may be used if the RTN-induced fluctuation is not large enough. Actually the implementation can be flexible and the binary bits can be converted from other parameters which are affected by RTN. Other implementations based on the same theory are equivalent to the representative scheme.

A. Randomness Modeling

1) Autocorrelation: Let X be the state of a trap. $X = 0/1$ indicates the emitted/captured state. The probabilities of the two states in a stationary state are given by

$$P(X=1) = P_1 = \frac{\bar{\tau}_e}{\bar{\tau}_c + \bar{\tau}_e},\ P(X=0) = P_0 = \frac{\bar{\tau}_c}{\bar{\tau}_c + \bar{\tau}_e}. \qquad (11)$$

Considering two time points s and $s+t$, the transition probabilities of a trap, which are also the prediction probability $P(X_{s+t}|X_s)$, are given by [33]

$$\begin{aligned} P(X_{s+t} = 1|X_s = 1) &= P_{11} = P_1 + P_0 e^{-\frac{t}{\tau_0}} \\ P(X_{s+t} = 0|X_s = 1) &= P_{10} = P_0 - P_0 e^{-\frac{t}{\tau_0}} \\ P(X_{s+t} = 0|X_s = 0) &= P_{00} = P_0 + P_1 e^{-\frac{t}{\tau_0}} \\ P(X_{s+t} = 1|X_s = 0) &= P_{01} = P_1 - P_1 e^{-\frac{t}{\tau_0}} \qquad (12) \end{aligned}$$

where P_{ij} means the probability of ending at state j after an elapsed time t, when starting from state i. In practice, t is the sampling period $t_s = (1/f_s)$, where f_s is the sampling frequency. According to (12), when t_s is long enough (e.g., $t_s > 3\tau_0$), the prediction probabilities are close to the stationary state probabilities P_0 and P_1, which means that knowing the previous state does not provide useful information for predicting the next state.

Let ΔA be the RTN-induced fluctuation (e.g., ΔV_{th}). Without loss of generality, we assume that ΔA is zero-mean. The autocorrelation function of the RTN-induced fluctuation is expressed as

$$\begin{aligned} C(s, s+t_s) = E\big(X_s X_{s+t_s}\big) &= (\Delta A)^2 \Big(P_0^2 P_1 P_{11} - P_0 P_1^2 P_{10} \\ &\qquad - P_0^2 P_1 P_{01} + P_1^2 P_0 P_{00}\Big) \\ &= (\Delta A)^2 \frac{\tau_0}{\bar{\tau}_e + \bar{\tau}_c} e^{-\frac{t_s}{\tau_0}}. \qquad (13) \end{aligned}$$

The autocorrelation function only depends on the duration time t_s so it can be written as $C(t_s)$. The first-order autocorrelation coefficient is given by

$$\rho(t_s) = \frac{C(t_s)}{C(0)} = e^{-\frac{t_s}{\tau_0}}. \qquad (14)$$

Increasing t_s can decrease the autocorrelation, which is consistent with the prediction probabilities as shown in (12).

To ensure high randomness, t_s should be high enough. For example, if $t_s \geq 3\tau_0$, the autocorrelation coefficient is less than 5%. The autocorrelation may be partly eliminated by applying some postprocessing methods, so the sampling frequency can be higher.

According to (11), $P_0 = P_1$ if and only if $\bar{\tau}_c = \bar{\tau}_e$, which is impossible in actual devices. As a result, postprocessing is always required. In this paper, we will take the popular von Neumann corrector [24] as an example to derive the autocorrelation, bias, and bit rate. The von Neumann corrector outputs "0" or "1" if two successive input bits are "01" or "10," but discards "00" and "11." After applying the von Neumann corrector, the autocorrelation coefficient can be approximated by

$$\rho_{vN}(t_s) \approx \frac{1}{-\frac{2}{e^{-\frac{t_s}{\tau_0}}} + 4 + \frac{3}{2P_0P_1 - 1}}, \quad t_s > 1.5\tau_0. \tag{15}$$

The derivation is complicated so we put it in Appendix A. The autocorrelation after the von Neumann corrector is applied is less than the original value given by (14). For example, if $f_s = (1/3\tau_0)$, (15) gives an autocorrelation of about 2.5%.

Till now, by deriving the autocorrelation coefficient, we have proved that when using a proper sampling frequency, the autocorrelation can be quite small such that the next bit cannot be predicted when the previous bit is known. This proves the randomness of single trap-induced RTN. Equation (15) can be used to estimate the proper sampling frequency when the von Neumann corrector is used.

2) Bias: Let the probabilities of ones and zeros are $0.5 + b$ and $0.5 - b$, respectively, where b is the bias, that is

$$b = \frac{\bar{\tau}_e - \bar{\tau}_c}{2(\bar{\tau}_c + \bar{\tau}_e)}. \tag{16}$$

The bias after the von Neumann corrector is applied is expressed as

$$\begin{aligned} b_{vN} &= P(\text{output "1"}|\text{has output}) - 0.5 \\ &= \frac{P_1P_{10}}{P_1P_{10} + P_0P_{01}} - 0.5 \equiv 0. \end{aligned} \tag{17}$$

As can be seen, the bias is completely eliminated, regardless of the sampling frequency and the original bias.

3) Bit Rate: For the von Neumann corrector, the ratio of the output bit rate to the sampling frequency equals half of the probability of observing "10" or "01," which is given by

$$\begin{aligned} \frac{R_{vN}}{f_s} &= \frac{1}{2}(P_0P_{01} + P_1P_{10}) = \frac{\tau_0}{\bar{\tau}_c + \bar{\tau}_e}\left(1 - e^{-\frac{t_s}{\tau_0}}\right) \\ &= \left(\frac{1}{4} - b^2\right)(1 - \rho(t_s)) < \frac{1}{4} \end{aligned} \tag{18}$$

where R_{vN} is the bit rate of the von Neumann corrector. R_{vN} depends on both the original bias and the autocorrelation. In any case, the bit rate is less than (1/4) of the sampling frequency. $\bar{\tau}_c = \bar{\tau}_e$ (i.e., $b = 0$) yields the maximum bit rate. Applying the first-order Taylor expansion to (18) yields

$$R_{vN} < \frac{1}{4t_s}\left(1 - e^{-\frac{t_s}{\tau_0}}\right) \approx \frac{1}{4t_s}\frac{t_s}{\tau_0} = \frac{1}{4\tau_0}. \tag{19}$$

Equation (19) gives the maximum ideal bit rate that the von Neumann corrector can achieve. It is achieved only when $\bar{\tau}_c = \bar{\tau}_e$ and f_s is high enough. However, using a high f_s leads to a high autocorrelation, so the maximum ideal bit rate cannot be achieved in practice.

B. Randomness Source Selection

As mentioned in Section III-A, RTN has a big uncertainty and the time constants have a wide range. It is difficult to make a specific transistor behave as what we expect. A feasible solution is to select an adequate transistor from a large transistor array [5]. We should select a transistor with exactly one observable trap, small time constants, and significant ΔV_{th}. To derive the probability of finding an adequate transistor, we first need the probability density function (PDF) of τ_0.

According to (2), the PDF of $\bar{\tau}_c$ is expressed as

$$\text{pdf}(\bar{\tau}_c) = \frac{1}{\bar{\tau}_c \ln \frac{\tau_{c,\max}}{\tau_{c,\min}}}, \quad \tau_{c,\min} \leq \bar{\tau}_c \leq \tau_{c,\max} \tag{20}$$

where $\tau_{c,\max}$ and $\tau_{c,\min}$ are the maximum and minimum values of $\bar{\tau}_c$, respectively. Although $\bar{\tau}_e$ is generated according to (3) with a small randomness on m, for simplicity in deriving the model, here we assume that m is fixed so $\bar{\tau}_e$ is linear with $\bar{\tau}_c$. Consequently, τ_0 is also linear with $\bar{\tau}_c$

$$\tau_0 = \frac{10^m}{10^m + 1}\bar{\tau}_c = \beta\bar{\tau}_c. \tag{21}$$

Then the PDF of τ_0 is expressed as

$$\text{pdf}(\tau_0) = \frac{1}{\tau_0 \ln \frac{\tau_{0,\max}}{\tau_{0,\min}}}, \quad \tau_{0,\min} \leq \tau_0 \leq \tau_{0,\max} \tag{22}$$

where $\tau_{0,\max} = \beta\tau_{c,\max}$ and $\tau_{0,\min} = \beta\tau_{c,\min}$.

Assuming that we have N transistors in total, the probability of finding at least one transistor, such that it has exactly one trap, $\tau_0 \in [\tau_{0,\min}, \delta\tau_{0,\min}]$, and $\Delta V_{\text{th}} \geq \gamma\langle\Delta V_{\text{th}}\rangle$, is expressed as

$$P = 1 - \left(1 - \frac{\ln\delta}{\ln\frac{\tau_{0,\max}}{\tau_{0,\min}}}\langle N_t\rangle e^{-\langle N_t\rangle} e^{-\gamma}\right)^N \tag{23}$$

where N can be solved from (23) when P is given. For example, if $\delta = 2$, $(\tau_{0,\max}/\tau_{0,\min}) = 10^4$, $\gamma = 2$, and $\langle N_t\rangle = 0.8$, then using $N \geq 1884$ can ensure a probability of larger than 0.999.

C. Numerical Results

To verify the proposed randomness model of single trap-induced RTN, the TRNG scheme as shown in Fig. 5 is simulated using the 22 nm PTM [32]. The nMOSFET is affected by a single trap. We use a fixed RTN profile in this test: $N_t = 1$, $\tau_0 = 10\ \mu$s, and $\Delta V_{\text{th}} = 40$ mV. We will analyze the performance under different $\bar{\tau}_c$ and $\bar{\tau}_e$ [(10) is always satisfied]. We adopt the concept of the approximate entropy (ApEn) [34] to evaluate the generated random numbers with the von Neumann corrector applied. All the results reported in this section are the mean values of five runs.

Fig. 6 shows the simulated ApEn, under different $(\bar{\tau}_e/\bar{\tau}_c)$ and f_s. Note that the maximum ideal ApEn is ln(2) $\approx$ 0.69315. As can be seen, ApEn is higher than 0.69 only when $f_s = 20$ and 45 kHz. Increasing f_s greatly decreases ApEn.

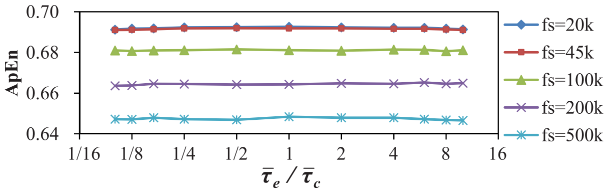


Fig. 6. ApEn of random numbers generated from single trap-induced RTN.

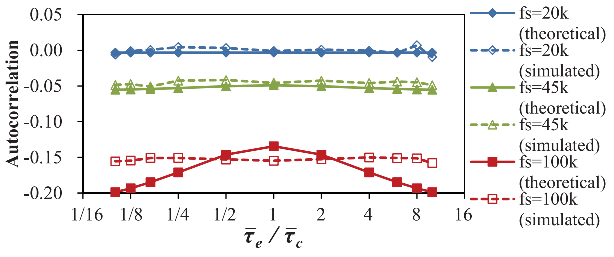


Fig. 7. Simulated and theoretical autocorrelation coefficients.

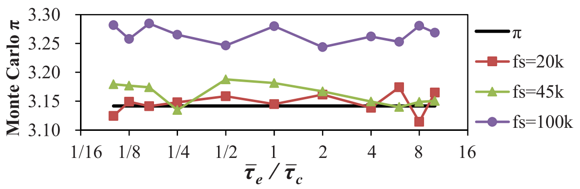


Fig. 8. Evaluated π by the MC method.

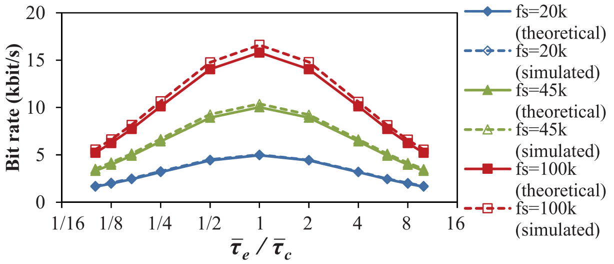


Fig. 9. Output bit rates of the von Neumann corrector.

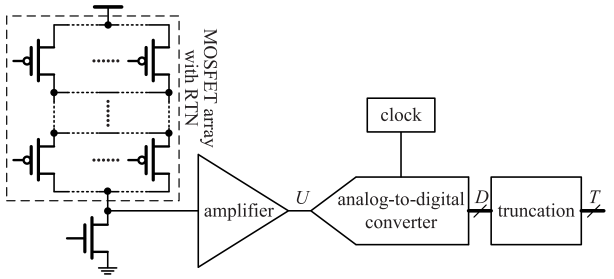


Fig. 10. Representative scheme for generating random numbers from multiple traps-induced RTN.

ApEn almost keeps constant when $(\bar{\tau}_e/\bar{\tau}_c)$ varies, revealing that the randomness is mainly determined by τ_0 and f_s. The simulated biases (not shown) are at the magnitude from 10^{-5} to 10^{-3} under different $(\bar{\tau}_e/\bar{\tau}_c)$, which means that the output bits are well balanced after applying the von Neumann corrector.

Fig. 7 shows the theoretical and simulated autocorrelation coefficients. The theoretical autocorrelation is predicted by (15). The simulated results are well consistent with the predictions when $f_s = 20$ and 45 kHz. However, when $f_s > (1/1.5\tau_0)$ (e.g., $f_s = 100$ kHz), the approximation of (15) leads to some errors (see Appendix A for an explanation). The generated random numbers are used to evaluate the value of π by the Monte Carlo (MC) method, as shown in Fig. 8. Results obtained by $f_s = 20$ and 45 kHz are acceptable. However, when $f_s = 100$ kHz, the results have big errors. According to the simulated ApEn, autocorrelation and MC π values, $f_s = (1/2.2\tau_0)$ is the maximum acceptable sampling frequency, corresponding to an autocorrelation of about 5%.

Fig. 9 shows the simulated and theoretical bit rates. The theoretical bit rates are predicted by (18). The simulated bit rates are well consistent with the predictions. As predicted by (18), the bit rate is always less than $(f_s/4)$. The bit rate decreases significantly when $(\bar{\tau}_e/\bar{\tau}_c)$ is far away from 1.0.

D. Summary

In this section, we have analyzed the autocorrelation coefficient, bias, and bit rate of a representative TRNG scheme based on single trap-induced RTN. As predicted by the proposed model and verified by the numerical results, using a too high sampling frequency greatly decreases the randomness of the output bits. In practice, the maximum sampling frequency is approximately $(1/2.2\tau_0)$, and the corresponding output bit rate is about $(0.4/(\bar{\tau}_c + \bar{\tau}_e))$. Selecting a trap with almost equal $\bar{\tau}_c$ and $\bar{\tau}_e$ maximizes the output bit rate. A large transistor array should be constructed to ensure that we can always find an adequate transistor to act as the randomness source.

V. Modeling Multiple Traps-Induced RTN

In this section, we first derive a theoretical randomness model for multiple traps-induced RTN, and then analyze the performance of a representative TRNG scheme based on multiple traps-induced RTN.

A simple idea to generate random numbers from multiple traps-induced RTN is to combine several individual circuits as shown in Fig. 5 in parallel such that multiple bits can be generated in one sampling. However, these are two practical problems. First, the time constants of many traps have a wide range so it is difficult to synchronize all the circuits using a unified sampling frequency. Second, the number of observable traps in each transistor is random, so some transistors do not show any fluctuation and some may show more-than-two-level fluctuations. In the single trap case, the two problems do not appear because we can select an adequate transistor from a large transistor array. Consequently, the effects of all the individual traps should be combined together to act as a single randomness source. Statistical laws and stochastic process theories will ensure that the overall RTN effect of numerous traps obeys a certain statistical rule.

Fig. 10 shows a representative scheme for generating random numbers from multiple traps-induced RTN. A number of transistors in which each is affected by multiple traps-induced RTN make up a transistor array. The transistor array can be regarded as a variable resistance affected by RTN. Although the fluctuation caused by each individual trap has only two discrete levels, the superposed fluctuation will have many discrete levels so it will look like a continuous signal. The superposed fluctuation is amplified and converted to digital words. The converted digital words have bias and autocorrelation, so the von Neumann corrector may also be applied, leading to a low bit rate. Considering the fact that in the converted digital words, high-order bits change slow and low-order bits change fast, low-order bits trend to be more random. This is a special feature of TRNGs based on multiple traps-induced RTN. In this section, we will investigate that by truncating a few high-order bits from the digital words, the remaining bits will have near-zero autocorrelation and bias. We will also show that the bit truncation scheme has a higher bit rate than the von Neumann corrector for TRNGs based on multiple traps-induced RTN.

A. Superposition of Multiple Lorentzian PSDs

We will first derive the PSD caused by multiple traps-induced RTN based on the statistical RTN model in Section III-A. The PSD of multiple traps-induced RTN is a superposition of PSDs of all the individual traps. By substituting (21) into (9), the Lorentzian PSD can be rewritten into the following form, with two random variables (ΔA and τ_0):

$$S(f) = 4\beta(1-\beta)(\Delta A)^2 \frac{\tau_0}{1+(2\pi f \tau_0)^2}. \tag{24}$$

Assuming that there are N traps in total, the superposition of all the individual PSDs is expressed as

$$S_N(f) = \sum_{i=1}^{N} 4\beta(1-\beta)(\Delta A_i)^2 \frac{\tau_{0,i}}{1+(2\pi f \tau_{0,i})^2} \tag{25}$$

where ΔA_i is the RTN-induced fluctuation. Note that each trap has its own amplitude and time constants which is mentioned in Section III-B. Actually, N is also a random variable. Since the summation of multiple independent Poisson distributions is still a Poisson distribution, N also obeys a Poisson distribution. For a large N, the summation in (25) can be converted into an integral

$$\begin{aligned} S_N(f) = 4\beta(1-\beta) \times \int_0^{\infty} (\Delta A)^2 \mathrm{pdf}(\Delta A)\mathrm{d}\Delta A \\ \times \int_{\tau_{0,\min}}^{\tau_{0,\max}} \frac{\tau_0}{1+(2\pi f \tau_0)^2}\mathrm{pdf}(\tau_0)\mathrm{d}\tau_0 \end{aligned} \tag{26}$$

where pdf(ΔA) depends on the implementation. The integral of $\int_0^{\infty} (\Delta A)^2 \mathrm{pdf}(\Delta A)\mathrm{d}\Delta A$ is always a constant (denoted as A) regardless of the detailed pdf(ΔA). Consequently, (26) can be converted into a closed form

$$\begin{aligned} S_N(f) &= \frac{4\beta(1-\beta)A}{\ln\frac{\tau_{0,\max}}{\tau_{0,\min}}} \int_{\tau_{0,\min}}^{\tau_{0,\max}} \frac{1}{1+(2\pi f \tau_0)^2}\mathrm{d}\tau_0 \\ &= \frac{2\beta(1-\beta)A}{\pi f \ln\frac{\tau_{0,\max}}{\tau_{0,\min}}}\big[\arctan(2\pi f \tau_{0,\max}) \\ &\qquad - \arctan\big(2\pi f \tau_{0,\min}\big)\big]. \end{aligned} \tag{27}$$

Equation (27) can be approximated by applying the first-order Taylor expansion to arctan according to the value of f [35]

$$S_N(f) \approx \begin{cases} \dfrac{\beta(1-\beta)A}{\pi^2 f^2 \ln\dfrac{\tau_{0,\max}}{\tau_{0,\min}}}\left(\dfrac{1}{\tau_{0,\min}} - \dfrac{1}{\tau_{0,\max}}\right), f \gg \dfrac{1}{2\pi\tau_{0,\min}} \\ \dfrac{\beta(1-\beta)A}{f\ln\dfrac{\tau_{0,\max}}{\tau_{0,\min}}}, \dfrac{1}{2\pi\tau_{0,\max}} \ll f \ll \dfrac{1}{2\pi\tau_{0,\min}} \\ \dfrac{4\beta(1-\beta)A\big(\tau_{0,\max}-\tau_{0,\min}\big)}{\ln\dfrac{\tau_{0,\max}}{\tau_{0,\min}}}, f \ll \dfrac{1}{2\pi\tau_{0,\max}}. \end{cases} \tag{28}$$

Considering that N is also a random variable, the superposed PSD has exactly the same form as (28) with the only difference on the amplitude A, since (28) is independent with N. For convenience, we still use (28) to express the superposed PSD. The superposed PSD shows three different shaped spectrums [i.e., white, $(1/f)$, and $(1/f^2)$]. Among them, the $(1/f)$ spectrum occupies a wide frequency range if $\tau_{0,\max} \gg \tau_{0,\min}$.

B. Randomness Modeling

1) Autocorrelation: The autocorrelation function can be calculated from the PSD based on the Wiener–Khinchin theorem [36], that is

$$C(t) = \int_0^{\infty} S_N(f)\cos(2\pi f t)\mathrm{d}f. \tag{29}$$

By applying the autocorrelation function of the band-limited $(1/f)$ spectrum [37] and ignoring the $(1/f^2)$ part which is very small, the autocorrelation function is approximated by

$$C(t) \approx \frac{A\beta(1-\beta)}{\ln\frac{\tau_{0,\max}}{\tau_{0,\min}}}\left(\frac{2}{\pi} + \ln\frac{\tau_{0,\max}}{t}\right), \tau_{0,\min} \le t \le \tau_{0,\max}. \tag{30}$$

We also have

$$C(0) = \int_0^{\infty} S_N(f)\mathrm{d}f = \beta(1-\beta)A\left(\frac{4}{\pi\ln\frac{\tau_{0,\max}}{\tau_{0,\min}}} + 1\right). \tag{31}$$

Then the autocorrelation coefficient is expressed as

$$\rho(t_s) = \frac{C(t_s)}{C(0)} \approx \frac{1}{2} + \frac{\ln\frac{\sqrt{\tau_{0,\min}\tau_{0,\max}}}{t_s}}{\frac{4}{\pi} + \ln\frac{\tau_{0,\max}}{\tau_{0,\min}}}, \tau_{0,\min} \le t_s \le \tau_{0,\max}. \tag{32}$$

When t_s is close to $\tau_{0,\min}$, the autocorrelation is close to 1.0 (if $\tau_{0,\max} \gg \tau_{0,\min}$). Although (32) is not applicable for $t_s < \tau_{0,\min}$, it is no doubt that the autocorrelation is approximately 1.0 in this case, because the sampling is faster than the fastest trap and two successive samplings tend to get identical values. To make the autocorrelation small, t_s should be close to $\tau_{0,\max}$, which means that the sampling frequency should match the slowest trap, leading to a very low bit rate. We will show that, after truncating a few high-order bits from the converted digital words, the autocorrelation of the remaining bits can be significantly eliminated and the bias is also close to zero, such that t_s is not required to be close to $\tau_{0,\max}$.

Let U, D, and T be the amplified fluctuation, the converted digital words, and the words after truncation, respectively. According to the central limit theorem, the overall RTN effect caused by numerous traps can be modeled by a Gaussian process. As a result, the fluctuation U follows a normal distribution. Let $\phi_U(\mu_U, \sigma_U; x)$ be the PDF of U, where μ_U and σ_U are the mean value and the standard deviation, respectively. Amplification does not affect the autocorrelation, so the autocorrelation coefficient of U equals $\rho(t_s)$ which is given by (32). U is converted to digital words D via an analog-to-digital converter (ADC) which encodes n bits (i.e., the output range is from 0 to $2^n - 1$). The ratio of $(\bar{\tau}_e/\bar{\tau}_c)$ can affect μ_U and σ_U. However, for a theoretical analysis, we assume that the full-scale range of the ADC matches $\mu_U \pm 3\sigma_U$. Actually, this can be achieved by adjusting the amplifier. We ignore the fluctuation out of the $\mu_U \pm 3\sigma_U$ range since its probability is negligible (i.e., 0.0027). The autocorrelation coefficient of D is also $\rho(t_s)$. To eliminate the autocorrelation of D, $n-k$ high-order bits of D are truncated and k low-order bits are kept. In what follows, we will derive the autocorrelation coefficient, bias, and bit rate of T.

First, we have the following probability for D:

$$\begin{aligned} P(D = i) &= P(iQ - 3\sigma_U \le X \le (i+1)Q - 3\sigma_U) \\ &= \int_{iQ-3\sigma_U}^{(i+1)Q-3\sigma_U} \phi_U(\mu_U, \sigma_U; x)\mathrm{d}x, 0 \le i \le 2^n - 1 \end{aligned} \tag{33}$$

where $Q = (6\sigma_U/2^n)$ is the quantized level of the ADC. Since $T = i$ if and only if the digital value expressed by the k low-order bits of D equals i, the probability of observing $T = i$ is expressed as

$$P(T = i) = \sum_{z=0}^{2^{n-k}-1} P\Big(D = i + z2^k\Big), 0 \le i \le 2^k - 1. \tag{34}$$

Since the integral of a normal distribution PDF has no analytical solution, (34) can be only estimated by numerical methods. We have found that if $n - k \ge 2$, all the $P(T = i)$'s are almost equal, that is

$$P(T = i) \approx \frac{1}{2^k}, 0 \le i \le 2^k - 1, k \le n - 2. \tag{35}$$

The maximum relative error of this approximation is less than 0.65%. An explanation of (35) is provided in Appendix B.

Considering two successive sampling time points s and $s+t_s$, the joint probability of observing $D_s = i$ and $D_{s+t_s} = j$ is given by

$$\begin{aligned} &P(D_s = i \cap D_{s+t_s} = j) \\ &\quad = \int_{iQ-3\sigma_U}^{(i+1)Q-3\sigma_U} \int_{jQ-3\sigma_U}^{(j+1)Q-3\sigma_U} \phi_2(\mu_U, \sigma_U, \rho_U(t_s); x, y)\mathrm{d}x\mathrm{d}y \\ &\quad 0 \le i, j \le 2^n - 1 \end{aligned} \tag{36}$$

where $\phi_2(\mu_U, \sigma_U, \rho_U(t_s); x, y)$ is the joint PDF of a bivariate normal distribution with an autocorrelation coefficient $\rho_U(t_s)$ which is given by (32). The joint probability of observing $T_s = i$ and $T_{s+t_s} = j$ is given by

$$\begin{aligned} &P(T_s = i \cap T_{s+t_s} = j) \\ &\quad = \sum_{z_1=0}^{2^{n-k}-1} \sum_{z_2=0}^{2^{n-k}-1} P\Big(D_s = i + z_1 2^k \cap D_{s+t_s} = j + z_2 2^k\Big) \\ &\quad 0 \le i, j \le 2^k - 1. \end{aligned} \tag{37}$$

The autocorrelation coefficient of T is expressed as

$$\rho_T(t_s) = \frac{\sum_{i=0}^{2^k-1} \sum_{j=0}^{2^k-1} ijP\big(T_s = i \cap T_{s+t_s} = j\big) - \mu_T^2}{\sigma_T^2} \tag{38}$$

where μ_T and σ_T are the mean value and standard deviation of T, which can be easily calculated based on (35).

Since (36) involves a double integral of the joint PDF of a bivariate normal distribution, (38) has no closed form. Equation (38) can be estimated by numerical methods, and then k can be selected such that $\rho_T(t_s)$ is small enough. We have derived a heuristic and effective method to directly calculate the optimal k such that $\rho_T(t_s)$ is negligible. The derivation is complicated so we put it in Appendix C. The optimal k is given by the following closed form:

$$k = \left\lfloor n - \log_2 \frac{6}{\sqrt{(\rho_U(t_s) - 1)\ln\left(2\pi\epsilon\sqrt{1 - \rho_U^2(t_s)}\right)}} \right\rfloor \tag{39}$$

where ϵ is a near-zero threshold to control the accuracy. We use $\epsilon = 10^{-6}$ in this paper. According to the explanation in Appendix C, if k is selected based on (39), $\rho_T(t_s) \approx 0$.

2) Bias: The bias of T is expressed as

$$b_T = \frac{1}{k} \sum_{i=0}^{2^k-1} \mathrm{ones}(i)P(T = i) - \frac{1}{2} \tag{40}$$

where ones(i) is the number of ones in the binary representation of integer i. Based on that the maximum relative error of the approximation of (35) is less than 0.65% (if $k \le n - 2$), we have

$$|b_T| < \frac{1}{k} \sum_{i=0}^{2^k-1} \left[\mathrm{ones}(i) \times \frac{0.0065}{2^k}\right] = 0.00325. \tag{41}$$

Equation (41) reveals that, although D is not uniformly distributed, the bias of T is close to zero after truncation.

3) Bit Rate: The output bit rate after truncation is simply expressed as kf_s. According to (32) and (39), increasing f_s will decrease k. f_s can be increased by 10×, 100×, etc., while k is decreased by only a few bits. Consequently, to maximize the bit rate, f_s should be as high as possible and close to $(1/\tau_{0,\min})$. If the von Neumann corrector is applied instead of bit truncation, the bit rate is lower than $(1/4)nf_s$. Calculation based on (39) reveals that, even if $\rho_U(t_s) = 0.95$, $k < (n/4)$ holds only when $n < 4$, leading to an impractical result $k < 1$. This reveals that the bit truncation scheme has a higher bit rate than the von Neumann corrector in practice.

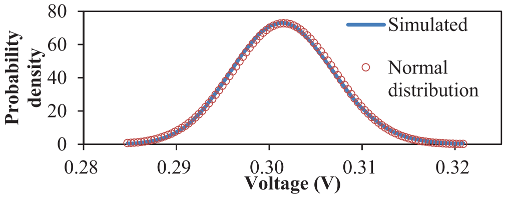


Fig. 11. Distribution sampled from the fluctuation caused by multiple traps-induced RTN, and the fitted normal distribution.

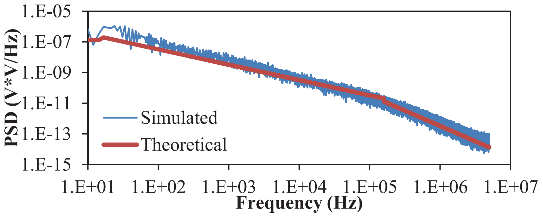


Fig. 12. Simulated and theoretical PSDs of multiple traps-induced RTN.

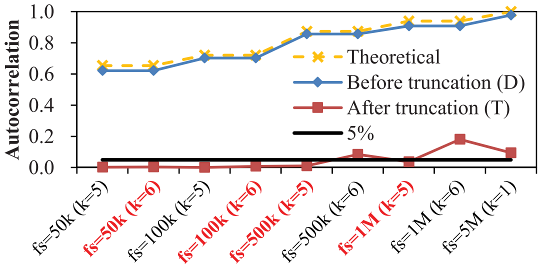


Fig. 13. Theoretical [by (32)] and simulated autocorrelation coefficients (before and after truncation).

C. Numerical Results

The representative TRNG scheme as shown in Fig. 10 with a 5 × 10 transistor array is simulated. The 22 nm PTM [32] is used. The widths of all the transistors are 50 nm. V_g of all the pMOSFETs and the nMOSFET are 0 and 0.8 V, respectively. We use the following RTN parameters for pMOSFETs: $\langle N_t \rangle = 2$, $\langle \Delta V_{\text{th}} \rangle = 20$ mV, $\tau_{0,\max} = 10$ ms, and $\tau_{0,\min} = 1\mu$ s. The RTN profile is randomly generated according to these parameters. The resolution of the ADC is $n = 8$. Since the purpose of this test is to verify the proposed randomness model, we assume that the amplifier can be adjusted such that $\mu_U \pm 3\sigma_U$ matches the full-scale range of the ADC. As a result, the ratio of $(\bar{\tau}_e/\bar{\tau}_c)$ which can affect μ_U and σ_U will have little impact on the final output. We use $\bar{\tau}_e = \bar{\tau}_c$ in this experiment. V_d of the nMOSFET is utilized as the randomness source. Fig. 11 shows an example (different runs give different examples) of the simulated distribution and the fitted normal distribution of V_d of the nMOSFET, in which the simulated distribution is converted from an 100-bin histogram. The fluctuation cased by multiple traps-induced RTN shows a nearly perfect normal distribution. Fig. 12 shows the simulated and theoretical PSDs of V_d of the nMOSFET. The theoretical PSD is calculated by (28), where $A \approx 1.202 \times 10^{-4}\text{V}^2$ is calculated from simulation. The simulated PSD is generally consistent with the theoretical PSD with a small difference. The difference mainly comes from the approximation of (28).

TABLE I
SIMULATION RESULTS OF THE TRNG BASED ON MULTIPLE TRAPS-INDUCED RTN

	ApEn	MC π	Bias	Bit rate (Mbit/s)
$f_s = 50\text{k}\ (k = 5)$	0.6926	3.1393	7.44E-04	0.25
$\boldsymbol{f_s = 50\text{k}\ (k = 6)}$	0.6927	3.1376	8.47E-04	0.3
$f_s = 100\text{k}\ (k = 5)$	0.6929	3.1358	4.37E-04	0.5
$\boldsymbol{f_s = 100\text{k}\ (k = 6)}$	0.6929	3.1425	4.73E-04	0.6
$\boldsymbol{f_s = 500\text{k}\ (k = 5)}$	0.6930	3.1344	3.96E-04	2.5
$f_s = 500\text{k}\ (k = 6)$	0.6924	3.1318	4.63E-04	3.0
$\boldsymbol{f_s = 1\text{M}\ (k = 5)}$	0.6927	3.1337	4.22E-04	5.0
$f_s = 1\text{M}\ (k = 6)$	0.6896	3.1299	5.07E-04	–
$f_s = 5\text{M}\ (k = 1)$	0.6864	3.0508	4.40E-04	–

Fig. 13 shows the theoretical and simulated autocorrelation coefficients, under different sampling frequencies and k. The theoretical autocorrelation is predicted by (32). The simulated autocorrelation before truncation is consistent with the predictions. The bit truncation scheme significantly eliminates the autocorrelation. The optimal k values estimated by (39) are marked as bold and red on the labels of the x-axis. When $f_s \leq (1/\tau_{0,\min})$, the optimal k ensures that the autocorrelation after truncation is less than 5%. However, when $f_s = 5$ MHz which is larger than $(1/\tau_{0,\min})$, the autocorrelation is larger than 5% even if only 1 bit is kept.

Table I shows the simulated ApEn, MC π values, bias, and bit rate. The ApEn results further reveal that the optimal k values calculated by (39) are correct. For the MC π value, a significant error is observed when $f_s = 5$ MHz due to the high autocorrelation. The biases of these cases are all at the magnitude of 10^{-4}, revealing that the bit truncation scheme can ensure a near-zero bias. According to these results, it can be concluded that when $f_s \leq (1/\tau_{0,\min})$, we can select an optimal k by (39) such that the autocorrelation is small enough to generate high-quality random numbers.

For the bit rate, clearly, among all of these cases, the maximum bit rate such that the randomness is guaranteed is achieved when $f_s = 1$ MHz and $k = 5$, which gives a bit rate of 5 Mbit/s. In this case, if the von Neumann corrector is applied instead of bit truncation, the bit rate is 1.92 Mbit/s, which is much lower than that of bit truncation.

D. Summary

In this section, we have derived a theoretical randomness model for multiple-traps induced RTN. We have demonstrated an interesting conclusion. When generating random numbers from multiple traps-induced RTN, the sampling frequency can be close to the switching frequency of the fastest trap. The high autocorrelation can be almost completely eliminated by truncating a few high-order bits from the converted digital words. The bias after truncation is also close to zero. We have provided a closed form to decide the optimal truncation.

VI. CASE STUDY: OSCILLATOR-BASED TRNG

In this section, we study an RO-based TRNG scheme and present how to determine key parameters for this TRNG, based on the proposed randomness models.

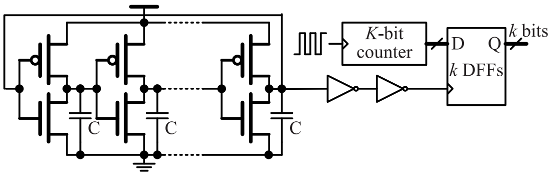


Fig. 14. RO-based TRNG based on multiple traps-induced RTN.

TABLE II
PARAMETERS USED IN THE RO-BASED TRNG

	22nm	32nm
V_{dd}	0.8V	0.9V
V_{th} (nMOSFET/pMOSFET)	0.503V/-0.461V	0.494V/-0.492V
$\tau_{0,max}/\tau_{0,min}$	10ms/1μs	10ms/1μs
$\langle N_t \rangle$ (nMOSFET/pMOSFET)	1.0/2.0	0.7/1.6
$\langle \Delta V_{th} \rangle$ [2], [28]	20mV	15mV

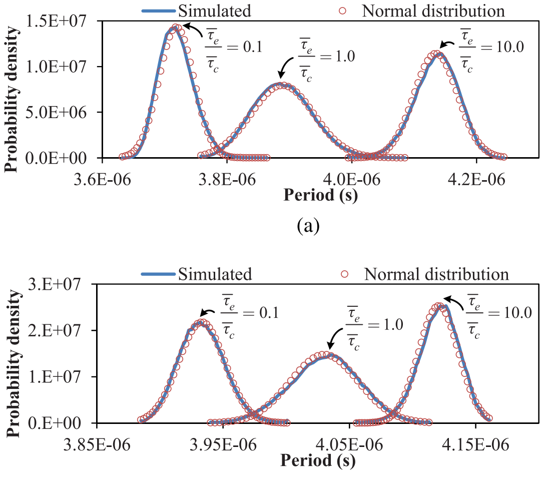
A. Overview

The TRNG scheme which contains a 25-stage inverter-based RO is shown in Fig. 14. Each stage has a load capacitance. Due to the RTN effect, the RO period will be random. A K-bit counter is used to count the rising edges of a high-frequency clock. The output of the RO clocks k ($k \leq K$) D-type flip-flops (DFF). The inputs of the k DFFs are connected to the k low-order bits of the counter, and the $K - k$ high-order bits of the counter are discarded. Clearly, the digital output of the counter is sampled at the end of each period of the RO output, so the RO period is converted to digital numbers. Due to the randomness in the RO period, the output of the k DFFs is also random. Actually, the theory of this scheme is the same as the representative TRNG based on multiple traps-induced RTN. The RO period is the randomness source which is affected by multiple traps-induced RTN, and the counter and the DFFs can be regarded as an ADC. We will explain how to determine k for this scheme in the next section.

B. Numerical Results

We test the RO-based TRNG at two technology nodes using the 22 and 32 nm PTM [32]. Parameters used in this test are listed in Table II. The widths of nMOSFETs and pMOSFETs are 80 and 40 nm, respectively. The load capacitance of each stage of the RO is 5 pF. The counter runs at 2 GHz. The width of the counter K is 20. Please note that K is not equivalent to the resolution of the ADC n as shown in Fig. 10. We will show how to calculate the equivalent n for this scheme in the following.

Fig. 15 shows the simulated distributions of the RO period and the fitted normal distributions, under different $(\bar{\tau}_e/\bar{\tau}_c)$. The RO period shows approximate normal distributions. The ratio of $(\bar{\tau}_e/\bar{\tau}_c)$ can affect the mean value and variance of the distribution. Since $(\bar{\tau}_e/\bar{\tau}_c)$ depends on many low-level factors and the manufacture process, it is difficult to know the exact $(\bar{\tau}_e/\bar{\tau}_c)$ by theoretical analysis. We consider three different values of $(\bar{\tau}_e/\bar{\tau}_c)$ in this test. The standard deviations of the three cases at the 22 nm node are 28, 50, and 35 ns, respectively. For the 32 nm node, they are 18, 27, and 16 ns, respectively. Obviously, $\bar{\tau}_c = \bar{\tau}_e$ yields the maximum variance of the RO period.



(a)

(b)

Fig. 15. Simulated distributions of the RO period, and the fitted normal distributions. (a) 22 nm. (b) 32 nm.

TABLE III
SIMULATION RESULTS OF THE RO-BASED TRNG

	22nm			32nm		
$\bar{\tau}_e/\bar{\tau}_c$	1.0	10.0	0.1	1.0	10.0	0.1
Autocorrelation	0.062%	0.015%	-0.25%	-0.055%	-0.087%	-0.23%
Bias	9.6E-05	6.1E-05	7.6E-05	-5.6E-05	-5.7E-05	-9.1E-04
ApEn	0.6930	0.6930	0.6930	0.6930	0.6930	0.6929
MC π	3.1311	3.1411	3.1376	3.1400	3.1543	3.1512
k	7	6	6	6	5	5
Bit rate (Mbit/s)	1.80	1.45	1.61	1.49	1.21	1.28

Now we show how to calculate the equivalent n and k in the RO-based TRNG, based on the proposed randomness model for multiple traps-induced RTN. The equivalent n is determined such that $[0, 2^n - 1]$ can cover most (i.e., 99.73%) of the digital outputs before truncation. Take the case of $(\bar{\tau}_e/\bar{\tau}_c) = 0.1$ at the 22 nm node as an example. When the counter runs at 2 GHz, $n \approx \log_2((28 \times 10^{-9}) \times (2 \times 10^9) \times 6) = 8.4$. The RO period is equivalent to the sampling period. According to (32), the autocorrelation of the RO period is about 80%. According to (39), we get $k = 6$, which means that we need six DFFs to sample the output of the counter. k values calculated by (39) are shown in Table III.

Table III shows the simulation results at the two technology nodes under different $(\bar{\tau}_e/\bar{\tau}_c)$ ratios. These results reveal that the generated random numbers of these cases are all of high quality and good randomness. The bit rate at the 32 nm node is lower than at the 22 nm node, mainly due to the smaller variance in the RO period at the 32 nm node.

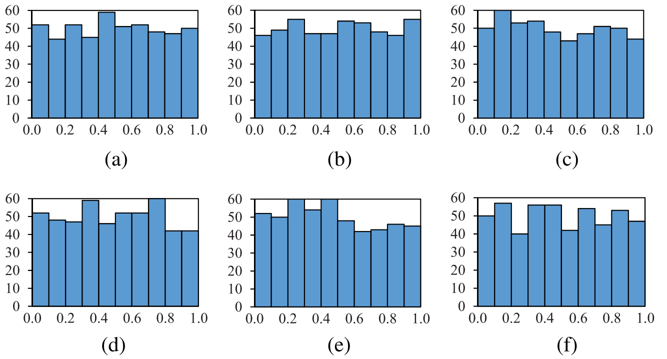
C. Randomness Test

The NIST test suite [23] is adopted to evaluate the randomness of the generated random numbers. For each case, we generate 500 random bit sequences by 500 independent simulations (each simulates a time interval of 50 ms), and then feed them into the NIST test suite. Table IV lists the pass rates of the reported p-values. A p-value larger than 0.01 indicates that

TABLE IV
PASS RATES (%) OBTAINED FROM 500 RUNS OF NIST

	22nm			32nm		
$\bar{\tau}_e/\bar{\tau}_c$	1.0	10.0	0.1	1.0	10.0	0.1
ApproximateEntropy	98.4	99.4	99.4	99.8	99.2	98.4
BlockFrequency	98.8	98.8	99.4	99.4	97.8	98.2
CumulativeSums[a]	98.9	98.8	98.8	98.9	98.4	98.2
FFT	98.4	98.4	98.8	99.2	98.2	98.4
Frequency	98.6	98.8	99.2	99.0	98.8	98.8
LinearComplexity	98.4	97.2	98.0	98.8	97.8	99.6
LongestRun	99.2	99.2	99.2	99.6	98.8	98.8
NonOverlappingTemplate[a]	98.8	98.9	98.8	98.9	98.8	98.8
OverlappingTemplate	99.0	98.6	99.4	98.6	98.8	98.4
RandomExcursions[a]	98.0	97.9	97.5	98.0	97.5	97.8
RandomExcursionsVariant[a]	98.6	97.9	98.4	98.5	98.8	98.5
Rank	99.6	99.4	99.0	98.8	98.4	99.0
Runs	99.0	98.4	98.4	98.6	98.8	99.0
Serial[a]	98.8	98.8	98.8	99.0	98.2	98.0
Universal	Insufficient length of bit sequence					

[a] These tests report multiple p-values, so each reported pass rate is the mean value of all the p-values from the corresponding test.



(a) (b) (c)

(d) (e) (f)

Fig. 16. Histograms of p-values obtained from 500 runs of the ApproximateEntropy test in NIST (the x-axis is the p-value and the y-axis is the count). (a) $\bar{\tau}_e/\bar{\tau}_c = 1.0$ (22 nm). (b) $\bar{\tau}_e/\bar{\tau}_c = 10.0$ (22 nm). (c) $\bar{\tau}_e/\bar{\tau}_c = 0.1$ (22 nm). (d) $\bar{\tau}_e/\bar{\tau}_c = 1.0$ (32 nm). (e) $\bar{\tau}_e/\bar{\tau}_c = 10.0$ (32 nm). (f) $\bar{\tau}_e/\bar{\tau}_c = 0.1$ (32 nm).

the test is passed. High pass rates are observed from Table IV, indicating good randomness of the generated random numbers. The distribution of p-values can also be utilized to evaluate the randomness of random numbers [23]. In theory, the distribution should be uniform. Fig. 16 shows histograms of p-values obtained from the ApproximateEntropy test (the block length is 5) in the NIST test suite. The six subfigures show approximate uniform distributions, indicating high randomness of the generated random numbers, as well.

VII. COMPARISON

In conventional noise-based TRNGs [16], [17], device noises are periodically sampled, amplified, and compared with a reference voltage to generate random bits. Since the thermal noise and $(1/f)$ noise are both tiny (typical magnitudes are from nV to μV) [38], strong amplifiers are required. On the contrary, RTN offers significant fluctuations in advanced technologies. Measured data have shown that ΔV_{th} caused by a single trap can be larger than 70 mV at the 22 nm node [1]. Such a large fluctuation can be easily converted to digital bits without suffering from variations, e.g., signal coupling problems.

TABLE V
SUMMARY OF THE PROPOSED RANDOMNESS MODELS

	TRNGs based on	
	single trap-induced RTN	multiple traps-induced RTN
Sampling frequency (f_s)	$\leq \frac{1}{2.2\tau_0}$	$\leq \frac{1}{\tau_{0,\min}}$
Post-processing	von Neumann corrector	Bit truncation (use Eq. (39) to determine the truncation)
Autocorrelation	$<$ 5% (by Eq. (15))	$\approx$ 0 (explained in Appendix C)
Bias	Exactly 0 (by Eq. (17))	$<$ 0.00325 (by Eq. (41))
Bit rate	$\frac{f_s\tau_0}{\bar{\tau}_c+\bar{\tau}_e}\left(1-e^{-\frac{t_s}{\tau_0}}\right) < \frac{f_s}{4}$ (by Eq. (18); $\bar{\tau}_c = \bar{\tau}_e$ maximizes the bit rate)	kf_s (select f_s close to $\frac{1}{\tau_{0,min}}$ to maximize the bit rate)

Conventional jitter-based TRNGs [19] typically use a slow and jittery clock to sample a fast clock. Due to the jitter of the slow clock, the fast clock is sampled at random positions so that random bits are generated. To achieve high randomness, the jitter must be larger than the period of the fast clock. However, measured data have shown that the jitter-to-mean period is at the magnitude of only 10^{-4} at the 0.18μm node [18], which is very small. Actually, the variation of the RO period cased by RTN can also be regarded as a "jitter." As shown in Fig. 15, the jitter-to-mean period caused by RTN is at the magnitude of 10^{-2}, such a big jitter allows that multiple bits can be generated from one sampling, resulting in a higher bit rate.

In summary, compared with conventional noise- and jitter-based TRNGs, the advantages of RTN-based TRNGs mainly come from the large fluctuations in advanced technologies. In addition, several studies have shown the increasing RTN effect due to the shrinking of the feature size [15], [27], [39], so RTN is becoming more significant. In practice, the RO jitter should be caused by all possible randomness sources, including RTN, $(1/f)$ noise, and white noise. However, measured data have shown that at the 22 nm node, RTN is the major noise source and much more important than $(1/f)$ noise [2]. This conclusion reveals that the RO jitter at advanced technology nodes is mainly caused by RTN.

VIII. CONCLUSION

In this paper, we have derived fundamental randomness models for RTN-based TRNGs. We have given theoretical models for the autocorrelation coefficient, bias, and bit rate of TRNGs based on both single trap- and multiple traps-induced RTN. We have given theoretical methodologies to determine key parameters for designing RTN-based TRNGs, such as the sampling frequency and the number of truncated bits. Table V briefly summarizes the most important points of the proposed randomness models. The proposed models have been verified by numerical simulations. An RO-based TRNG at two technology nodes is studied based on the model of multiple traps-induced RTN. The proposed randomness models will be verified by fabricated chips in our future work.

APPENDIX A
DERIVATION OF (15)

Let $\{X_n\}$ be the sampled binary sequence, and $\{Y_n\}$ be an intermediate sequence

$$Y_n = \begin{cases} X_{2n}, & \text{if } X_{2n} \neq X_{2n+1} \\ 2, & \text{if } X_{2n} = X_{2n+1}. \end{cases} \tag{42}$$

If all the 2s in $\{Y_n\}$ are discarded, we get a new sequence $\{Z_n\}$, which is the output of the von Neumann corrector for $\{X_n\}$. When $f_s < (1/1.5\tau_0)$, the second-order autocorrelation coefficient between X_n and X_{n+2} is less than 5%, so $\{X_n\}$ can be approximately regarded as a Markov chain, and thus, we have the following probabilities for $\{Y_n\}$:

$$\begin{aligned}
&P(Y_{n+1} = 1 \cap Y_n = 1) \approx P_1 P_{10} P_{01} P_{10} \triangleq y_{11} \\
&P(Y_{n+1} = 2 \cap Y_n = 1) \approx P_1 P_{10}(P_{00}P_{00} + P_{01}P_{11}) \triangleq y_{21} \\
&P(Y_{n+1} = 1 | Y_n = 2) \approx \frac{P_{10}(P_0 P_{00} P_{01} + P_1 P_{11} P_{11})}{P_0 P_{00} + P_1 P_{11}} \triangleq y_{1|2} \\
&P(Y_{n+1} = 2 | Y_n = 2) \\
&\quad \approx \frac{P_0 P_{00}(P_{00}P_{00} + P_{01}P_{11}) + P_1 P_{11}(P_{10}P_{00} + P_{11}P_{11})}{P_0 P_{00} + P_1 P_{11}} \\
&\quad \triangleq y_{2|2}.
\end{aligned} \tag{43}$$

Clearly, "11" in $\{Z_n\}$ is generated from "11," "121," "1221," $\cdots$ in $\{Y_n\}$. However, the probabilities of observing "11" in $\{Z_n\}$ and observing "11," "121," "1221," $\cdots$ in $\{Y_n\}$ are different, because the lengths of $\{Y_n\}$ and $\{Z_n\}$ are different. The difference of the probabilities equals the ratio of their lengths, which can be obtained from (18). Consequently, we have

$$\begin{aligned}
&P(Z_{n+1} = 1 \cap Z_n = 1) \\
&\quad \approx \frac{f_s}{2R_{vN}}(y_{11} + y_{1|2}y_{21} + y_{1|2}y_{2|2}y_{21} + y_{1|2}y_{2|2}y_{2|2}y_{21} + \cdots) \\
&\quad = \frac{f_s}{2R_{vN}}\left(y_{11} + \frac{y_{1|2}y_{21}}{1 - y_{2|2}}\right).
\end{aligned} \tag{44}$$

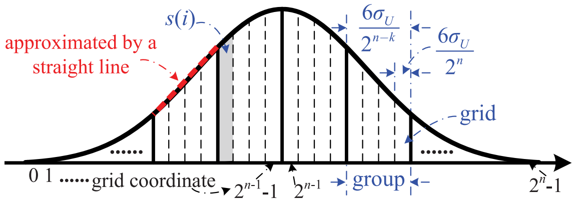
According to (17), the probabilities of ones and zeros in $\{Z_n\}$ are balanced, and thus, the mean value and the variance of $\{Z_n\}$ are 0.5 and 0.25, respectively. The autocorrelation coefficient of $\{Z_n\}$ is expressed as

$$\begin{aligned}
&\rho_{vN}(t_s) \\
&\quad = \frac{P(Z_{n+1} = 1 \cap Z_n = 1) - 0.5 \times 0.5}{0.25} \\
&\quad \approx 2\left[1 - \frac{1 - 2P_0P_1 + 3P_0P_1 e^{-\frac{t_s}{\tau_0}} - P_0P_1 e^{-\frac{3t_s}{\tau_0}}}{2 - 4P_0P_1 - e^{-\frac{t_s}{\tau_0}}(1 - 8P_0P_1) + e^{-\frac{2t_s}{\tau_0}}(1 - 4P_0P_1)}\right] - 1.
\end{aligned} \tag{45}$$

The terms with respect to $e^{-(2t_s/\tau_0)}$ and $e^{-(3t_s/\tau_0)}$ can be ignored since they are quite small, and then we can simply get (15).

APPENDIX B
EXPLANATION OF (35)

We first consider (34). The physical meaning of (34) is illustrated in Fig. 17. The 2-D region constructed by the PDF of the normal distribution $\phi_U(\mu_U, \sigma_U; x)$ and the interval $[\mu_U - 3\sigma_U, \mu_U + 3\sigma_U]$ of the x-axis is divided into 2^{n-k} groups. All the groups have the same length $(6\sigma_U/2^{n-k})$ on the x-axis. Each group is further divided into 2^k grids. All the grids have the same length $(6\sigma_U/2^n)$ on the x-axis. Each grid has a coordinate from 0 to $2^n - 1$. Let $s(i)$ be the area of grid i. It is clear that $s(i) = P(D = i)$, so we have



$s(i)$
approximated by a straight line
$\frac{6\sigma_U}{2^{n-k}}$
$\frac{6\sigma_U}{2^n}$
grid
0 1 ...... grid coordinate
2^{n-1}-1
2^{n-1}
group
2^n-1

Fig. 17. Illustration of (34) and the derivation of (35).

$$P(T = i) = \sum_{z=0}^{2^{n-k}-1} s\left(i + z2^k\right), 0 \leq i \leq 2^k - 1. \tag{46}$$

Due to the symmetry of the PDF, we have

$$\begin{aligned}
P(T = i) = &\sum_{z=0}^{2^{n-k-1}-1} \left[s\left(i + z2^k\right) + s\left((z+1)2^k - 1 - i\right)\right] \\
&0 \leq i \leq 2^k - 1.
\end{aligned} \tag{47}$$

It is easy to check that

$$P(T = i) \equiv P\left(T = 2^k - 1 - i\right), 0 \leq i \leq 2^{k-1} - 1. \tag{48}$$

So we only need to consider the difference between half of all the $P(T = i)$'s. The difference between $P(T = i)$ and $P(T = j)$ is expressed as

$$\begin{aligned}
&P(T = i) - P(T = j) \\
&\quad = \sum_{z=0}^{2^{n-k-1}-1} \begin{bmatrix} s\left(i + z2^k\right) - s\left(j + z2^k\right) \\ + s\left((z+1)2^k - 1 - i\right) \\ - s\left((z+1)2^k - 1 - j\right) \end{bmatrix} \\
&\quad 0 \leq i \neq j \leq 2^{k-1} - 1.
\end{aligned} \tag{49}$$

The four area terms on the right side of (49) are in the same group for the same z. If $n - k$ is big enough, the length of each group will be small enough, such that the PDF curve within each group can be approximated by a straight line with little error, as shown in Fig. 17. If the approximation has no error, the algebraic sum of the four area terms is exactly 0, and thus, $P(T = i) = P(T = j)$ [for $0 \leq i, j \leq 2^k - 1$]. Of course, the PDF curve within each group is not an exact straight line, so we have $P(T = i) \approx P(T = j)$ in practice. Since the integral of a normal distribution PDF has no closed form, numerical experiments have verified that when $n - k = 2$, the maximum relative error of the approximation as shown in (35) is about 0.65%. According to the above explanation, increasing $n - k$ (i.e., decreasing k) makes the approximation more accurate so the error is smaller.

APPENDIX C
HEURISTIC DERIVATION OF (39)

We first consider the physical meaning of (37). Like the 2-D case shown in Appendix B, the 3-D region

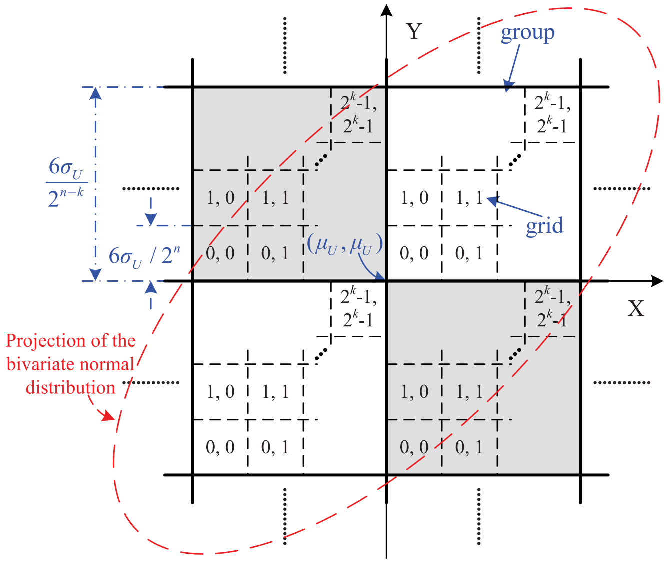


Fig. 18. Illustration of (37) and the derivation of (39).

constructed by the PDF of the bivariate normal distribution $\phi_2(\mu_U, \sigma_U, \rho_U(t_s); x, y)$ and the square region $[\mu_U - 3\sigma_U, \mu_U + 3\sigma_U] \times [\mu_U - 3\sigma_U, \mu_U + 3\sigma_U]$ on the *xy* plane is divided into $2^{n-k} \times 2^{n-k}$ groups. The bottom of each group is a square and the length of one side is $(6\sigma_U/2^{n-k})$. Fig. 18 shows the four nearest groups to the central point (μ_U, μ_U). Each group is further divided into $2^k \times 2^k$ grids. The bottom of each grid is also a square and the length of one side is $(6\sigma_U/2^n)$. Each grid has a local coordinate (i, j) $(0 \leq i, j \leq 2^k - 1)$ in its group. If the contour of the PDF $\phi_2(\mu_U, \sigma_U, \rho_U(t_s); x, y)$ is projected onto the *xy* plane, it will show an ellipse, as shown in Fig. 18. The equation of the projected ellipse is given by

$$(x - \mu_U)^2 + (y - \mu_U)^2 - 2\rho_U(t_s)(x - \mu_U)(y - \mu_U) = -2\sigma_U^2\left(1 - \rho_U^2(t_s)\right)\ln\left(2\pi\epsilon\sqrt{1 - \rho_U^2(t_s)}\right) \quad (50)$$

where ϵ is the relative height of the contour. ϵ should be near zero to get accurate results. A bigger $\rho_U(t_s)$ leads to a higher eccentricity of the ellipse, i.e., the ellipse looks narrower. Clearly, $P(T_s = i \cap T_{s+t_s} = j)$ (37) equals the total volume of all the grids whose local coordinates are (i, j) $(0 \leq i, j \leq 2^k - 1)$. We can ignore all the grids out of the ellipse, since the cumulative probability out of the ellipse is quite small if $\epsilon \approx 0$.

Like the 2-D case explained in Appendix B, the PDF surface within each grid can be approximated by a plane, such that all $P(T_s = i \cap T_{s+t_s} = j)$'s are almost equal with an exception when the ellipse cannot cover sufficient grids. It can be analyzed from Fig. 18 that, if the ellipse cannot fully cover the two shaded groups, the volumes of the shaded grids in the ellipse cannot be balanced even if the approximation by planes is accurate enough, leading to a big difference between $P(T_s = i \cap T_{s+t_s} = j)$'s. If this happens, the autocorrelation coefficient of T will be high according to (38). Consequently, the optimal k to ensure a low autocorrelation coefficient should be selected such that the ellipse can just fully cover the two shaded groups, i.e., the two points $(\mu_U + (6\sigma_U/2^{n-k}), \mu_U - (6\sigma_U/2^{n-k}))$ and $(\mu_U - (6\sigma_U/2^{n-k}), \mu_U + (6\sigma_U/2^{n-k}))$ are both on the curve of the ellipse. Substituting either point into (50) will yield (39).

References

[1] N. Tega *et al.*, "Increasing threshold voltage variation due to random telegraph noise in FETs as gate lengths scale to 20 nm," in *Proc. VLSI Technol. Symp.*, Honolulu, HI, USA, Jun. 2009, pp. 50–51.

[2] N. Tega *et al.*, "Reduction of random telegraph noise in high-k/metal-gate stacks for 22 nm generation FETs," in *Proc. IEEE Int. Electron Devices Meeting (IEDM)*, Baltimore, MD, USA, Dec. 2009, pp. 1–4.

[3] M. Stipčević and Ç. K. Koç, "True random number generators," Dept. Elect. Comput. Eng., Univ. California Santa Barbara, Santa Barbara, CA, USA, Tech. Rep., 2012.

[4] N. Liu, N. Pinckney, S. Hanson, D. Sylvester, and D. Blaauw, "A true random number generator using time-dependent dielectric breakdown," in *Proc. VLSI Circuits (VLSIC) Symp.*, Honolulu, HI, USA, Jun. 2011, pp. 216–217.

[5] R. Brederlow, R. Prakash, C. Paulus, and R. Thewes, "A low-power true random number generator using random telegraph noise of single oxide-traps," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, San Francisco, CA, USA, Feb. 2006, pp. 1666–1675.

[6] M. J. Kirton and M. J. Uren, "Noise in solid-state microstructures: A new perspective on individual defects, interface states and low-frequency (1/f) noise," *Adv. Phys.*, vol. 38, no. 4, pp. 367–468, 1989.

[7] A. Hajimiri, S. Limotyrakis, and T. H. Lee, "Jitter and phase noise in ring oscillators," *IEEE J. Solid-State Circuits*, vol. 34, no. 6, pp. 790–804, Jun. 1999.

[8] A. A. Abidi, "Phase noise and jitter in CMOS ring oscillators," *IEEE J. Solid-State Circuits*, vol. 41, no. 8, pp. 1803–1816, Aug. 2006.

[9] U. Guler and G. Dundar, "Modeling CMOS ring oscillator performance as a randomness source," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 61, no. 3, pp. 712–724, Mar. 2014.

[10] N. Göv, M. K. Mihcak, and S. Ergun, "True random number generation via sampling from flat band-limited Gaussian processes," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 58, no. 5, pp. 1044–1051, May 2011.

[11] T. Nagumo, K. Takeuchi, T. Hase, and Y. Hayashi, "Statistical characterization of trap position, energy, amplitude and time constants by RTN measurement of multiple individual traps," in *Proc. IEEE Int. Electron Devices Meeting (IEDM)*, San Francisco, CA, USA, Dec. 2010, pp. 28.3.1–28.3.4.

[12] K. Abe, A. Teramoto, S. Sugawa, and T. Ohmi, "Understanding of traps causing random telegraph noise based on experimentally extracted time constants and amplitude," in *Proc. IEEE Int. Rel. Phys. Symp. (IRPS)*, Monterey, CA, USA, Apr. 2011, pp. 4A.4.1–4A.4.6.

[13] A. P. van der Wel, E. A. M. Klumperink, L. K. J. Vandamme, and B. Nauta, "Modeling random telegraph noise under switched bias conditions using cyclostationary RTS noise," *IEEE Trans. Electron Devices*, vol. 50, no. 5, pp. 1378–1384, May 2003.

[14] T. Nagumo, K. Takeuchi, S. Yokogawa, K. Imai, and Y. Hayashi, "New analysis methods for comprehensive understanding of random telegraph noise," in *Proc. IEEE Int. Electron Devices Meeting (IEDM)*, Baltimore, MD, USA, Dec. 2009, pp. 1–4.

[15] K. Fukuda, Y. Shimizu, K. Amemiya, M. Kamoshida, and C. Hu, "Random telegraph noise in flash memories—Model and technology scaling," in *Proc. IEEE Int. Electron Devices Meeting (IEDM)*, Washington, DC, USA, Dec. 2007, pp. 169–172.

[16] C. S. Petrie and J. A. Connelly, "A noise-based IC random number generator for applications in cryptography," *IEEE Trans. Circuits Syst. I, Fundam. Theory Appl.*, vol. 47, no. 5, pp. 615–621, May 2000.

[17] M. Matsumoto *et al.*, "1200 um2 physical random-number generators based on SiN MOSFET for secure smart-card application," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, Feb. 2008, pp. 414–624.

[18] M. Bucci, L. Germani, R. Luzzi, A. Trifiletti, and M. Varanonuovo, "A high-speed oscillator-based truly random number source for cryptographic applications on a smart card IC," *IEEE Trans. Comput.*, vol. 52, no. 4, pp. 403–409, Apr. 2003.

[19] G. K. Balachandran and R. E. Barnett, "A 440-nA true random number generator for passive RFID tags," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 55, no. 11, pp. 3723–3732, Dec. 2008.

[20] P. Z. Wieczorek and K. Golofit, "Dual-metastability time-competitive true random number generator," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 61, no. 1, pp. 134–145, Jan. 2014.

[21] S. N. Dhanuskodi, A. Vijayakumar, and S. Kundu, "A chaotic ring oscillator based random number generator," in *Proc. IEEE Int. Symp. Hardw. Oriented Security Trust (HOST)*, Arlington County, VA, USA, May 2014, pp. 160–165.

[22] C.-Y. Huang, W. C. Shen, Y.-H. Tseng, Y.-C. King, and C.-J. Lin, "A contact-resistive random-access-memory-based true random number generator," *IEEE Electron Device Lett.*, vol. 33, no. 8, pp. 1108–1110, Aug. 2012.

[23] A. Rukhin *et al.*, *A Statistical Test Suite for Random and Pseudorandom Number Generators for Cryptographic Applications*, document 800-22, Nat. Inst. Stand. Technol., Gaithersburg, MD, USA, 2010.

[24] J. von Neumann, "Various techniques used in connection with random digits," *Nat. Bureau Stand. Appl. Math. Series*, vol. 12, no. 3, pp. 36–38, 1951.

[25] J. P. Campbell *et al.*, "The origins of random telegraph noise in highly scaled SiON nMOSFETs," in *Proc. IEEE Int. Integr. Rel. Workshop Final Report*, South Lake Tahoe, CA, USA, Oct. 2008, pp. 105–109.

[26] T. Grasser *et al.*, "A unified perspective of RTN and BTI," in *Proc. IEEE Int. Rel. Phys. Symp.*, South Lake Tahoe, CA, USA, Jun. 2014, pp. 4A.5.1–4A.5.7.

[27] S. Realov and K. L. Shepard, "Analysis of random telegraph noise in 45-nm CMOS using on-chip characterization system," *IEEE Trans. Electron Devices*, vol. 60, no. 5, pp. 1716–1722, May 2013.

[28] N. Tega, "Study on impact of random telegraph noise on scaled MOSFETs," Ph.D. dissertation, Grad. School Pure Appl. Sci., Univ. Tsukuba, Tsukuba, Japan, 2014.

[29] M. Tanizawa *et al.*, "Application of a statistical compact model for random telegraph noise to scaled-SRAM Vmin analysis," in *Proc. VLSI Technol. Symp. (VLSIT)*, Honolulu, HI, USA, Jun. 2010, pp. 95–96.

[30] L. W. Nagel, "SPICE 2: A computer program to stimulate semiconductor circuits," Ph.D. dissertation, Dept. Electr. Eng. Comput. Sci., Univ. California, Berkeley, CA, USA, 1975.

[31] UC Berkeley Device Group. (2015). *BSIM4*. [Online]. Available: http://www-device.eecs.berkeley.edu/bsim/?page=BSIM4

[32] Nanoscale Integration and Modeling (NIMO) Group ASU. (2015). *Predictive Technology Model (PTM)*. [Online]. Available: http://ptm.asu.edu/

[33] R. da Silva and G. I. Wirth, "Logarithmic behavior of the degradation dynamics of metal-oxide-semiconductor devices," *J. Stat. Mech. Theory Exp.*, vol. 2010, no. 4, 2010, Art. ID P04025.

[34] S. M. Pincus, "Approximate entropy as a measure of system complexity," *Proc. Nat. Acad. Sci.*, vol. 88, no. 6, pp. 2297–2301, 1991.

[35] F. N. Hooge, T. G. M. Kleinpenning, and L. K. J. Vandamme, "Experimental studies on 1/f noise," *Rep. Progr. Phys.*, vol. 44, no. 5, pp. 479–532, 1981.

[36] C. Chatfield, *The Analysis of Time Series: An Introduction*, 6th ed. New York, NY, USA: Chapman and Hall, 2013.

[37] F. N. Hooge and P. A. Bobbert, "On the correlation function of 1/f noise," *Physica B: Condens. Mat.*, vol. 239, nos. 3–4, pp. 223–230, 1997.

[38] T. C. Carusone, D. Johns, and K. W. Martin, *Analog Integrated Circuit Design*, 2nd ed. Hoboken, NJ, USA: Wiley, 2011.

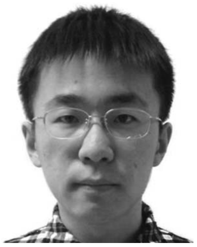
[39] A. Ghetti *et al.*, "Scaling trends for random telegraph noise in decananometer flash memories," in *Proc. IEEE Int. Electron Devices Meeting (IEDM)*, San Francisco, CA, USA, Dec. 2008, pp. 1–4.

Xiaoming Chen (S'12–M'15) received the B.S. and Ph.D. degrees from the Department of Electronic Engineering, Tsinghua University, Beijing, China, in 2009 and 2014, respectively.

Since 2014, he has been a Post-Doctoral Researcher of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA, USA. He is also a Part-Time Researcher with the Department of Electronic Engineering, Tsinghua University. His current research interests include hardware security and Internet of things.

Lin Wang received the B.S. degree from the School of Mathematical Sciences, Nankai University, Tianjin, China, in 2009, and the Ph.D. degree from the Academy of Mathematics and Systems Sciences, Chinese Academy of Sciences, Beijing, China, in 2014.

She is a Lecturer with the School of Statistics, Capital University of Economics and Business, Beijing. Her current research interests include probability and mathematical statistics, computational biology, and biostatistics.



Boxun Li (S'13) received the B.S. degree from the Department of Electronic Engineering, Tsinghua University, Beijing, China, in 2009, where he is currently pursuing the M.S. degree.

His current research interests include energy-efficient hardware computing system design and parallel computing based on GPUs.

Yu Wang (S'05–M'07–SM'14) received the B.S. and Ph.D. (Hons.) degrees from the Department of Electronic Engineering, Tsinghua University, Beijing, China, in 2002 and 2007, respectively.

He is currently an Associate Professor with the Department of Electronic Engineering, Tsinghua University. He has authored and co-authored over 130 papers in refereed journals and conferences. His current research interests include parallel circuit analysis, application specific hardware computing (especially on the brain-related problems), and power/reliability aware system design methodology.

Dr. Wang was a recipient of the IBM X10 Faculty Award in 2010, the Best Paper Award in IEEE Annual Symposium on Very Large Scale Integration 2012, the Best Poster Award in International Symposium on Highly-Efficient Accelerators and Reconfigurable Technologies 2012, and six best paper nominations in Asia and South Pacific Design Automation Conference (ASPDAC), International Conference on Hardware/Software Codesign and System Synthesis, and International Symposium on Low Power Electronics and Design (ISLPED).

Xin Li (S'01–M'06–SM'10) received the B.S. and M.S. degrees in electronics engineering from Fudan University, Shanghai, China, in 1998 and 2001, respectively, and the Ph.D. degree in electrical and computer engineering from Carnegie Mellon University, Pittsburgh, PA, USA, in 2005.

He is currently an Associate Professor with the Department of Electrical and Computer Engineering, Carnegie Mellon University. His current research interests include integrated circuit and signal processing.

Prof. Li was a recipient of the National Science Foundation Faculty Early Career Development Award in 2012, the IEEE Donald O. Pederson Best Paper Award in 2013, the Best Paper Award from the Design Automation Conference in 2010, two IEEE/ACM William J. McCalla International Conference on Computer Aided Design Best Paper Awards in 2004 and 2011, and the Best Paper Award from the International Symposium on Integrated Circuits in 2014.

Yongpan Liu (M'07–SM'15) received the B.S., M.S., and Ph.D. degrees from the Department of Electronic Engineering, Tsinghua University, Beijing, China, in 1999, 2002, and 2007, respectively.

He is an Associate Professor with the Department of Electronic Engineering, Tsinghua University. He has published over 60 papers and designed several sensor chips, including the first nonvolatile processor (THU1010N). His current research interests include low power design, emerging circuits and systems, and electronic design automation.

Prof. Liu was a recipient of the ISLPED2012/2013 Design Contest Award, and several best paper nominations. He served as a Technical Program Committee Member for Design Automation Conference, ASPDAC, and Asian Solid-State Circuits Conference.

Huazhong Yang (M'97–SM'00) received the B.S. degree in microelectronics and the M.S. and Ph.D. degrees in electronic engineering from Tsinghua University, Beijing, China, in 1989, 1993, and 1998, respectively.

In 1993, he joined the Department of Electronic Engineering, Tsinghua University, where he is a specially appointed Professor of the Cheung Kong Scholars Program. He has authored and co-authored over 300 technical papers and 70 granted patents. His current research interests include wireless sensor networks, data converters, parallel circuit simulation algorithms, nonvolatile processors, and energy-harvesting circuits.