Analysis of Random Telegraph Noise in 45-nm CMOS Using On-Chip Characterization System

Simeon Realov, *Member, IEEE*, and Kenneth L. Shepard, *Fellow, IEEE*

***Abstract*—An on-chip variability characterization system implemented in a 45-nm CMOS process is used for direct time-domain measurements of random telegraph noise (RTN) in small-area devices. A procedure for automated extraction of RTN parameters from large volumes of measured data is developed. Statistics for number of traps, N_T, and single-trap amplitudes, ΔV_{th}, are studied across device polarity, bias, and gate area. A Poisson distribution is used to model N_T and a log-normal distribution is used to model ΔV_{th}. The scaling of the two statistics across gate dimensions is discussed; the expected value of N_T is shown to scale with $(L - \Delta L)^{-1}$, whereas the expected value of ΔV_{th} is shown to scale with $W^{-1}(L - \Delta L)^{-0.5}$. The two statistics are combined in a compact RTN probabilistic model representing the statistics of the overall ΔV_{th} fluctuations because of RTN. This model is demonstrated to give accurate predictions of the tails of the measured RTN distributions at the 95th percentile level, which scale with $W^{-1}(L - \Delta L)^{-1.5}$. A comparison between nMOS and pMOS devices shows that pMOS devices exhibit both a higher average number of traps and a larger average single-trap ΔV_{th} amplitude, leading to a comparatively larger overall impact of RTN.**

***Index Terms*—45-nm, characterization, CMOS, on-chip, random telegraph noise (RTN), scaling, statistics.**

I. Introduction

RANDOM telegraph noise (RTN) is a low-frequency noise phenomenon in semiconductor devices, which is manifested as discrete random jumps in the drain current amplitude of a field effect transistor because of the capture and emission of charge carriers in potential traps at the channel/oxide interface, as shown in Fig. 1. RTN is an issue of growing concern in advanced CMOS technology nodes, especially as minimum channel length scales down to 45-nm and below [1]–[5].

Much effort is directed toward the characterization and statistical modeling of RTN. Measured statistical distributions of RTN amplitude are skewed and exhibit fat tails, which makes RTN a potential major source of performance variability in high-density circuit blocks, such as static random access memories [1]. Better modeling of the statistical distribution of drain current amplitude fluctuations because of RTN and its scaling as a function of device dimensions is needed to be able to anticipate and properly address design issues arising from this new source of device variability.

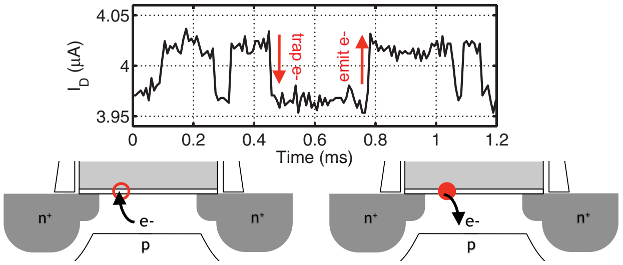

Fig. 1. Measured two-level RTN waveform along with an illustration of the underlying carrier trapping process.

This paper describes a comprehensive methodology for the characterization and analysis of time-domain RTN measurements, as well as the development of a robust statistical model for the prediction of the impact of RTN on device performance. Section II introduces an on-chip CMOS characterization system implemented in a 45-nm low-power process used for time-domain RTN characterization of large sample sets of small-area devices [6]. An automated methodology for the extraction of RTN parameters from time-domain measurements, as needed for the analysis of a large volume of measurement data, is developed in Section III. In Section IV, the statistics of number of traps and single-trap amplitudes are analyzed across device polarity, bias, and geometry and are combined in a compact statistical model for the prediction of the overall impact of RTN on device performance. In particular, the use of a log-normal single-trap amplitude distribution is shown to yield a statistical model that accurately captures the tails of the measured RTN distributions at the 95th percentile level. Section V concludes.

II. Measurement Setup

A simplified top-level schematic of the on-chip characterization system is shown in Fig. 2. The system consists of three major blocks: 1) an on-chip switching matrix, used to individually address transistors from the device-under-test (DUT) array; 2) a four-channel digital-to-analog converter (DAC), used to supply each of the four DUT terminal bias voltages; and 3) a measurement unit, which consists of a current- and voltage-mode analog-to-digital converters (ADCs), used to perform accurate on-chip current–voltage (I–V) characterization. The biasing DAC is a resistor-string DAC with four independent channels and an 8-bit resolution operating with

Manuscript received February 8, 2012; accepted February 22, 2013. Date of current version April 18, 2013. The review of this paper was arranged by Editor J. D. Cressler.

The authors are with the Department of Electrical Engineering, Columbia University, New York, NY 10027 USA (e-mail: srealov@ee.columbia.edu; shepard@ee.columbia.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/TED.2013.2254118

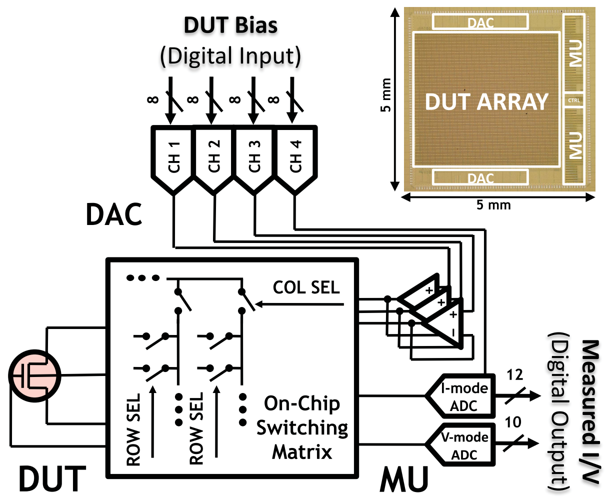

Fig. 2. Simplified top-level schematic of the on-chip characterization system and its implementation in a 45-nm low-power bulk CMOS process (inset).

a $V_{\text{LSB}} = 4.3$ mV. The current- and voltage-mode ADCs are based on a dual-slope integrator topology. The current-mode ADC has a 12-bit resolution with $I_{\text{LSB}} = 2.44$ nA; the voltage-mode ADC has 10-bit resolution with $V_{\text{LSB}} = 0.976$ mV. The measurement sampling rates are optimized with respect to the input signal strength and vary between 125 and 75 kHz. Two copies of the system are integrated into the same chip to allow simultaneous nMOS and pMOS device characterization for increased measurement throughput. For the duration of all measurements, the temperature of the test chip is maintained at a steady 25 °C using an external thermal control system.

The gate-to-source bias range, V_{GS}, covered by the RTN measurements consists of five equally spaced bias points, starting at 0.63 V and ending at 0.80 V, with a drain-to-source potential, V_{DS}, set to 50 mV. This bias range is chosen such that the device operates in the strong-inversion linear region, where

$$I_D = \mu C'_{\text{ox}} \frac{W}{L} \left(V_{\text{GS}} - V_{\text{th}} - \frac{V_{\text{DS}}}{2} \right) V_{\text{DS}} \tag{1}$$

gives a reasonable small-signal approximation for the relationship between I_{D} and V_{th}.

The sample size at each bias point is 2^{21} sample points, which results in a measurement duration between 17 and 28 s per bias point. Although RTN traps can have characteristic capture and emission times of up to a few hours or more [7], characterizing a large number of devices at such long-time intervals for a number of bias points and across a number of different geometries is impractical. In addition, when a subset of the data is sampled at intervals up to four times longer appreciable differences in the final results are not observed, implying that the vast majority of traps present in the population are detectable using the shorter sampling interval.

The measurement DUT sample set consists of an orthogonal set of minimum-length devices ($L = 0.04$ μm, $W = 0.2, 0.4, 0.6$ μm) and minimum-width devices ($W = 0.2$ μm, $L = 0.04, 0.8, 0.11$ μm), enabling the study of RTN properties as a function of device geometry by capturing the effect of varying both W and L. Both nMOS and pMOS devices are measured, where the bias for the two device polarities is set such that $V_{\text{GS}_{\text{nMOS}}} = V_{SG_{\text{pMOS}}}$ and $V_{\text{DS}_{\text{nMOS}}} = V_{\text{SD}_{\text{pMOS}}}$. The DUTs are organized in statistical sets of 78 devices per chip, and a total of four chips are measured for an overall statistical sample set of 312 DUT per DUT type type. The size of the sample set is sufficiently large to enable observing statistics at the 95th percentile level.

III. Parameter Extraction

Because of the high volume and random nature of the measured data, a fully automated analysis methodology has to be developed to extract RTN parameters from $I_{\text{D}}(t)$ measurements. In particular, the quantities of interest that need to be extracted are the number of observed traps, N_T, and the RTN amplitude associated with individual traps.

A. Time Lag Plot

A time lag plot (TLP), also known as a lag scatter plot, is a tool for analyzing autocorrelation in time-series data and can be used to analyze time-domain RTN measurements [8]. TLPs are constructed by plotting data sampled at the ith time interval, t_i, versus data sampled at $ith+1$ time interval, t_{i+1}. As the $I_{\text{D}}(t)$ waveform lingers at different RTN levels, the measured data at t_i and t_{i+1} is similar, and RTN levels appear as data clusters along the $I_{\text{D}}(t_i) = I_{\text{D}}(t_{i+1})$ diagonal of the TLP. This approach, however, is impractical for analyzing large amounts of RTN data because of the lack of an unambiguous way to identify RTN levels, particularly in the presence of white and $1/f$ noise in the measured data.

To overcome these limitations, we use an enhanced TLP data analysis technique. The new approach aims to automate detection of individual RTN levels even in the presence of additional noise, making the analysis of large statistical data sets feasible. In this case, we record the frequency with which each point of the TLP is occupied, transforming the TLP into a two-dimensional histogram of $I_{\text{D}}(t_i)$ versus $I_{\text{D}}(t_{i+1})$ with a bin size equal to one I_{LSB}. A comparison between using a standard and an enhanced TLP is shown in Fig. 3. Analyzing the frequency with which a data point from the lag scatter plot is occupied makes the detection of distinct RTN levels possible even in cases where the amount of noise present makes the standard TLP approach impractical. These enhanced TLPs make possible the extraction of the number of traps as well as the trap amplitudes from RTN waveforms.

B. Extraction of Number of Traps, N_T

The first step in the extraction of the number of active traps in a device, N_T, is the detection of the number of distinct RTN levels present in the measured signal. RTN levels are identified along the line $I_{\text{D}}(t_i) = I_{\text{D}}(t_{i+1})$, the diagonal of the enhanced TLP as shown in Fig. 3(c). The diagonal of the enhanced TLP also gives information about the frequency with which each point is occupied, and the number of local maxima extracted along the enhanced TLP diagonal represents the number of detected RTN levels, N_L. Local maxima are generally well-defined, because of the large number of samples in each RTN

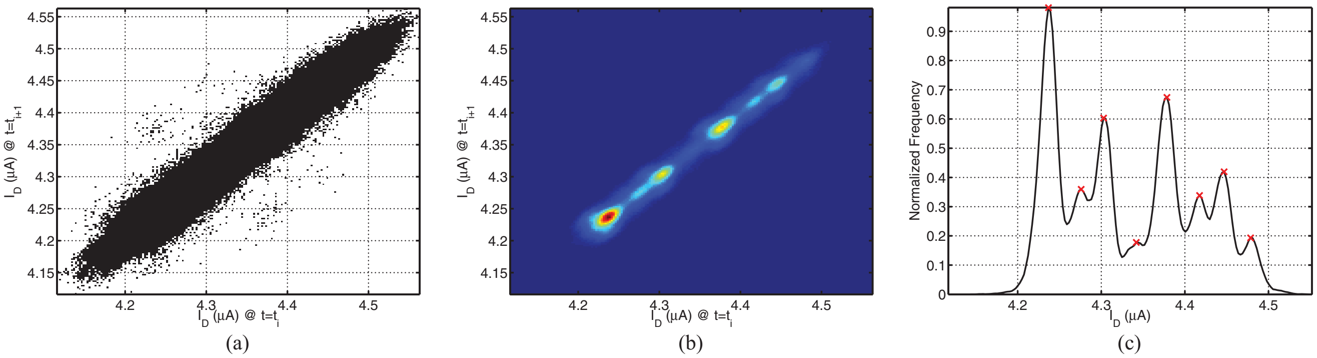

Fig. 3. Comparison between (a) standard and (b) enhanced TLP for a triple-trap RTN signal. (c) Enhanced TLP diagonal along with detected RTN levels.

waveform, and can be easily extracted using a threshold-based peak detection algorithm.

To extract the number of RTN traps present, the following relationship is used [8]:

$$N_T = \mathrm{ceil}(\log_2(N_L)) \tag{2}$$

where ceil(x) is the ceiling function, which rounds up x to the nearest integer value. Equation 2 is based on the assumption that the effects of multiple traps are additive. For example, the superposition of two traps results in four RTN levels and the superposition of three traps results in eight RTN levels. The ceil(x) function is used to account that not all possible combinations of RTN trap occupancy may be expressed within the measurement time interval.

C. Extraction of Single-Trap Amplitude, ΔV_{th}

The first step in extracting single-trap RTN amplitudes from measured data is to extract the changes in the measured current, ΔI_{D}, due to a single trap. The most straight-forward approach is to extract ΔI_{D} from RTN measurements where only a single trap is observed [7]. As the two peaks in the enhanced TLP diagonal represent the two RTN levels in a single-trap RTN waveform, the distance between the two peaks gives ΔI_{D}.

One concern about this approach is that it limits the number of extracted single-trap amplitudes to measurements based only on single-trap RTN signals. This can be problematic, especially if single-trap waveforms are rare in the studied population. To extract a larger number of single-trap amplitudes, the enhanced TLP diagonal can be used to extract multiple ΔI_{D} measurements from multitrap RTN waveforms. In particular, as the effects of individual RTN traps are assumed to be additive, multitrap RTN signals are superpositions of multiple single-trap RTN signals. Each trap has its own characteristic capture and emission times, independent of any other traps in the same device, which can be used to distinguish RTN levels because of different traps. The relative heights of the peaks in the diagonal of the enhanced TLP give an indication of the relative frequency with which each trap is occupied. Therefore, the distance between the two highest peaks indicates the amplitude, $\Delta I_{\mathrm{D},1}$, associated with the dominant RTN trap; the distance between the highest peak and the third highest peak indicates the amplitude, $\Delta I_{\mathrm{D},2}$, of the second dominant trap measured while the device is in its preferred state with regards to the first dominant trap.

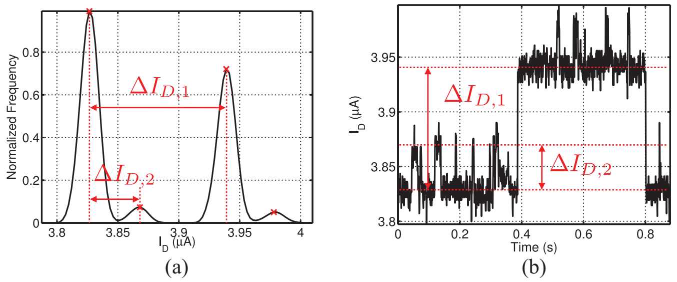

Fig. 4. Extraction of two independent ΔI_D measurements from a single-two-trap RTN waveform. (a) Measurements along the TLP diagonal. (b) The corresponding measured RTN waveform.

This concept is illustrated in the case of a two-trap RTN signal in Fig. 4. However, it applies to any multitrap RTN waveform. Although extending this approach to extract more than two single-trap amplitudes from multitrap signals becomes challenging, extracting just two individual RTN amplitudes from each multitrap RTN waveform provides ample data with which to study the statistics of individual trap amplitudes.

Once the single-trap ΔI_{D} amplitude is extracted, it can be expressed in terms of a ΔV_{th} variation. Modeling RTN as a ΔV_{th} effect is consistent with the identification of charge-carrier fluctuations as the main mechanism for the observed ΔI_{D} fluctuations [4], [7], [9]. According to (1)

$$\Delta I_{\mathrm{D}} = \mu C'_{\mathrm{ox}} \frac{W}{L} V_{\mathrm{DS}} \Delta V_{\mathrm{th}}. \tag{3}$$

To avoid any issues associated with IR drops across the DUT array, and more importantly, the dependence of the mobility, μ, on V_{GS}, 3 can be expressed as

$$\Delta V_{\mathrm{th}} = \frac{\Delta I_{\mathrm{D}}}{G_{\mathrm{M}}} \tag{4}$$

where G_{M} is given by

$$G_{\mathrm{M}} \equiv \frac{\partial I_{\mathrm{D}}}{\partial V_{\mathrm{GS}}} = \mu C'_{\mathrm{ox}} \frac{W}{L} V_{\mathrm{DS}}. \tag{5}$$

G_{M} can be measured directly for each individual DUT by performing an I-V sweep of I_{D} as a function of V_{GS} for a V_{DS} nominally set to 50 mV, and the values of G_{M} at the bias points where the RTN measurements are performed can be

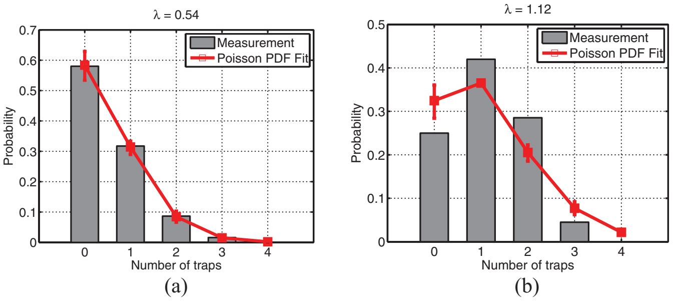

Fig. 5. Representative examples of measured PDFs for N_T and the corresponding Poisson fits for a nMOS and pMOS device with $W/L = 0.4/0.04$ μm at $|V_{GS}| = 0.72$ mV; in both cases the Poisson PDF (6) fits the measured data well. (a) nMOS, $\lambda = 0.54$. (b) pMOS, $\lambda = 1.12$.

extracted. Any variations of V_{DS} and μ as a function of V_{GS} cancel out in the extraction of ΔV_{th} using 4, as they affect the measured G_{M} and the measured ΔI_{D} proportionally.

IV. Statistical Modeling of RTN

To accurately model the overall amplitude variations in I_{D} because of RTN, a compact statistical model that encompasses the combined effects of amplitude variations and variations in the number of traps is needed. One approach is to separately model the statistics of the number of traps, N_T, and the statistics of single-trap amplitudes, ΔV_{th}, and then combine the two to construct a complex model for the predicting the overall variation in I_{D} [7]. This approach allows us to observe and model the two statistics independently as a function of device dimensions in an attempt to uncover the basic mechanisms driving the RTN scaling trends.

A. Statistics of Number of Traps, N_T

The statistics of N_T are widely reported in the literature to follow a Poisson probability distribution [7], [8], [10], [11]. Although generally no theoretical basis is given for this interpretation, intuitively, a Poisson distribution is well-suited for the modeling of discrete random events that occur within a fixed area with a given average rate and independently of one another. As such, the Poisson distribution should lend itself well to the modeling of the random occurrence of potential traps along the channel/oxide interface of FETs.

The probability density function of the Poisson distribution expressed in the context of predicting N_T is given by

$$f_T(N_T;\lambda) = \frac{\lambda^{N_T} e^{-\lambda}}{N_T!} \qquad (6)$$

where λ is the population mean of N_T. λ is the only parameter describing the Poisson distribution and is, therefore, the parameter of interest to be extracted from the measured distributions of N_T.

Fig. 5 shows Poisson fits of measured PDFs of N_T for an nMOS and a pMOS $W/L = 0.4/0.04$ μm device at mid-bias, with $|V_{\text{GS}}| = 0.72$ V. These fits are representative of the entire DUT bias and parameter space. As expected, the Poisson distribution gives an accurate representation of the statistics of N_T.

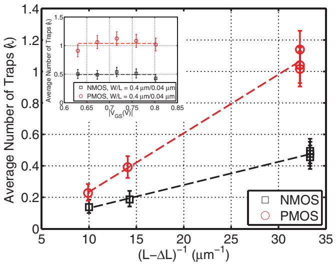

Fig. 6. Plot of the mean λ measured across bias versus the inverse of the effective channel length given by $1/(L\Delta L)$; λ remains constant across bias as shown in the inset for representative nMOS and pMOS device sample sets with $W/L = 0.4/0.04$ μm.

The scaling behavior of λ with device dimensions, W and L, is shown in Fig. 6 where the average λ across bias is plotted against the inverse of the effective gate length, given by $(L - \Delta L)^{-1}$, where ΔL is the overlap between the gate and the source/drain diffusions. An estimate for $\Delta L = 9$ nm is extracted from gate capacitance measurements using a technique similar to that described in [12], where careful extraction of the intrinsic gate capacitance as a function of the drawn dimensions, W and L, enables accurate estimation of the effective dimensions of the device; more details on our on-chip capacitance characterization methodology can be found in [13]. Across bias, λ remains relatively constant, as shown in the inset. Based on data from both nMOS and pMOS measurements, it appears that λ is largely independent of W and is inversely proportional to $L - \Delta L$. When comparing nMOS with pMOS devices, in every instance, the pMOS devices exhibit a higher average number of traps, which is consistent with earlier published results [8].

Theoretical analysis [7] predicts that the average number of traps should be proportional to the area under the gate, given by WL, rather than scale inversely with the effective length. However, such analysis ignores the issue of observability, and in particular, RTN is not observed in large-area devices because of reduced single-trap amplitude and the tendency of multiple RTN signals to combine and form $1/f$ noise [14]. Therefore, while the actual number of traps present in a device can grow proportionally to the gate area, it is still possible for the average number of observed RTN traps, λ, to scale inversely with the effective channel length, as observed here.

B. Statistics of Single-Trap Amplitude Fluctuations, ΔV_{th}

Accurately characterizing the statistical distribution of single-trap RTN amplitude fluctuations is vital for constructing an accurate compact model for the prediction of overall ΔV_{th} fluctuations because of RTN. Although results in the literature, based on both device simulation and experimental measurements, indicate that the distribution of single-trap ΔV_{th} is skewed and exhibits a fat tail, there is disagreement on which distribution captures the statistical effects best. In particular, two distributions are considered: the exponential distribution [5], [7], [8], [15], given by

$$f_e(\Delta V_{\text{th}};\sigma_e) = \frac{1}{\sigma_e} e^{-\frac{\Delta V_{\text{th}}}{\sigma_e}} \qquad (7)$$

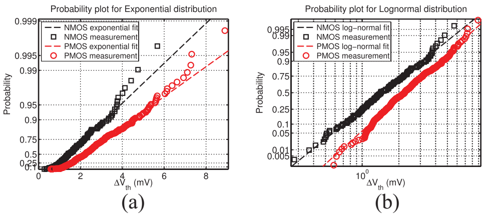

Fig. 7. Representative measured ΔV_{th} distributions for a $W/L = 0.4/0.04$ μm nMOS and pMOS device fitted using (a) exponential PDF and (b) log-normal PDF (dashed lines: ideal fits); in either case, a log-normal PDF more accurately models the tails of the measured distribution.

where σ_e is a parameter, which represents the population mean, and the log-normal distribution [11], [16], given by

$$f_l(\Delta V_{\text{th}}; V_{\text{th0}}, \sigma_l) = \frac{1}{\sigma_l \Delta V_{\text{th}} \sqrt{2\pi}} e^{-\frac{1}{2\sigma_l^2}(\ln \Delta V_{\text{th}} - \ln V_{\text{th0}})^2} \quad (8)$$

where σ_l is a dimensionless parameter representing the log-normal shape, and V_{th0} is given by

$$V_{\text{th0}} = e^{\mu} \quad (9)$$

with μ representing the mean of the distribution of $\ln(\Delta V_{\text{th}})$.

Although the exponential distribution is more commonly used for the modeling of the statistics of single-trap RTN amplitude, we find that the log-normal distribution yields a better fit to the measured data. Fig. 7 shows examples of representative single-trap ΔV_{th} distributions for an nMOS and a pMOS device fit to both an exponential and a log-normal distribution. Both distributions appear to offer acceptable fits to the measured data, but in either case, the log-normal distribution is better at modeling the tails of the observed statistical distribution. In particular, compared with the log-normal fit, the exponential fit over-predicts both the low end and the high end of the measured distribution.

Based on fits using the log-normal distribution, it is instructive to consider the expected value of ΔV_{th}, $E[\Delta V_{\text{th}}]$, given by

$$E[\Delta V_{\text{th}}] = V_{\text{th0}} e^{\sigma^2/2}. \quad (10)$$

A plot of extracted $E[\Delta V_{\text{th}}]$ averaged across bias is shown in Fig. 8. The inset in Fig. 8 shows that $E[\Delta V_{\text{th}}]$ is insensitive to gate bias conditions, similar to what is reported in [16]. pMOS devices exhibit higher single-trap amplitudes as compared with nMOS devices, which is also observed in [7]. In terms of area dependence, $E[\Delta V_{\text{th}}]$ is shown to be inversely proportional to $W(L-\Delta L)^{0.5}$. This functional dependence can be attributed to current crowding along the length of the device, as traps in that location have a greater overall effect on modulating the charge in the channel [7]. Such current crowding can be expected to have a similar effect to that of percolation paths in the onset of strong inversion as discussed in [5] where the same functional dependence is reported. The only exception are devices with $L = 0.11$ μm, but the small average number of observed traps in these devices makes it possible that the statistics of single-trap amplitudes are inaccurate simply because of the smaller available sample set.

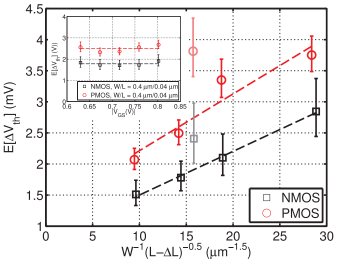

Fig. 8. Plot of the mean $E[\Delta V_{th}]$ averaged across bias versus W^{-1} $(L - \Delta L)^{-0.5}$, where $E[\Delta V_{th}]$ remains constant across bias as shown in the inset for $W/L = 0.4/0.04$ μm nMOS and pMOS devices; dashed lines: trend lines; faded data points correspond to devices with $L = 0.11$ μm, which do not follow the observed trend, possibly because of an insufficient number of samples.

C. Complex CDF Model for Overall RTN Fluctuations

The final step of constructing a statistical model for overall ΔV_{th} fluctuations (referred to, from this point on, as ΔV_{th}^* to distinguish them from single-trap ΔV_{th} fluctuations) is to combine the statistics of number of traps, N_T, with the statistics of single-trap amplitudes, ΔV_{th}, into one comprehensive statistical model. Assuming the effects of individual traps are additive, i.e., individual trapping and detrapping events are independent of one another and act in superposition, then the PDF for n observed traps can be derived using the successive convolution of n single-trap distributions [7]. Using the log-normal PDF to model the statistics of single-trap amplitude (8), we can express the PDF of a system of n traps as

$$f_{l,n}(\Delta V_{\text{th}}; V_{\text{th0}}, \sigma_l, n) = \int_{-\infty}^{\infty} f_{l,n}(\Delta V_{\text{th}} - u; V_{\text{th0}}, \sigma_l, n-1) \times f_l(u; V_{\text{th0}}, \sigma_l)\, \mathrm{d}u. \quad (11)$$

The relative contribution of an RTN system with n observed traps to the overall distribution of ΔV_{th}^* can be derived from the Poisson distribution of N_T given by 6 as

$$a_n = P(N_T = n) = \frac{\lambda^n e^{-\lambda}}{n!}. \quad (12)$$

Finally, the two statistics can be combined by multiplying each a_n coefficient by the corresponding $f_{l,n}(\Delta V_{\text{th}}; V_{\text{th0}}, \sigma_l, n)$; a delta function, $\delta_0(x)$, is used to represent the distribution of devices with no traps. The products are summed as n goes to infinity to give $f_c(\Delta V_{\text{th}}^*; V_{\text{th0}}, \sigma_l, \lambda)$, the overall PDF of ΔV_{th}^*, as

$$f_c(\Delta V_{\text{th}}^*; V_{\text{th0}}, \sigma_l, \lambda) = a_0 \delta_0(\Delta V_{\text{th}}^*) + \sum_{i=1}^{\infty} a_i f_{l,n}(\Delta V_{\text{th}}^*; V_{\text{th0}}, \sigma_l, i). \quad (13)$$

Equation 13 can be used to derive the cumulative distribution function (CDF) of ΔV_{th}^* as given by

$$F_c(\Delta V_{\text{th}}^*; V_{\text{th0}}, \sigma_l, \lambda) = \int_0^{\Delta V_{\text{th}}^*} f_c(x; V_{\text{th0}}, \sigma_l, \lambda)\, \mathrm{d}x. \quad (14)$$

Fig. 9 shows example CDFs derived using parameter values extracted from the measured distributions of N_T and ΔV_{th}, as described in Section IV-A and Section IV-B, respectively. To underscore the importance of using a log-normal distribution

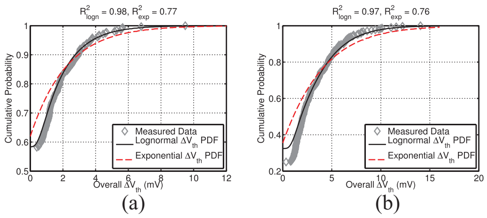

Fig. 9. Representative fits of measured overall ΔV^*_{th} CDFs (grey diamonds) for nMOS and pMOS devices with $W/L = 0.4/0.04$ μm; the plot shows a comparison between using solid black line: a log-normal ΔV_{th} PDF and dashed red line: an exponential ΔV_{th} PDF to construct the model CDF; R^2 values quoted for both cases above the individual graphs show that using a log-normal PDF yields a considerably better fit to the measured data. (a) nMOS. (b) pMOS.

to model ΔV_{th} (8), in contrast to an exponential one (7) [7], both functional forms are considered. The modeled CDFs are compared with the actual measured ΔV^*_{th} distributions. R^2 values are calculated to help evaluate the goodness of the fit, where R^2 is the coefficient of determination; the closer the value of R^2 is to 1, the better the fit.

In all cases, when a log-normal PDF is used to model the single-trap ΔV_{th} (solid black lines in Fig. 9), the fits are excellent, with a mean R^2 value of 0.97 across all samples. On the other hand, using an exponential PDF for ΔV_{th} (dashed red lines in Fig. 9) results in much poorer fits to the measured data, with a mean R^2 of 0.73 across all samples. This comparison once again demonstrates that the statistics of single-trap amplitude are better modeled by a log-normal distribution.

The measured CDFs of ΔV^*_{th} fit well the estimated CDFs calculated based on extracted parameters for N_T and ΔV_{th}, demonstrating that our compact statistical model gives an accurate representation of the statistical behavior of RTN and can be used to predict the total impact of RTN with high confidence even at the tails of the distribution. To further underscore this point, Fig. 10 shows a comparison between measured and predicted ΔV^*_{th} across the entire sample set at the 95th percentile level. In all cases, the agreement between prediction and measurement is excellent.

A scaling trend of the 95th percentile of ΔV^*_{th} proportional to $W^{-1}(L-\Delta L)^{-1.5}$ is observed. This can be traced to the scaling of the number of traps, N_T, with $(L-\Delta L)^{-1}$, as shown in Fig. 6, and the scaling of ΔV_{th} with W^{-1} $(L-\Delta L)^{-0.5}$, as shown in Fig. 8. The tail of the overall ΔV^*_{th} scales with $W^{-1}(L-\Delta L)^{-1.5}$, indicating that the comparative impact of RTN on V_{th} is expected to worsen with device scaling in relation to the effect of random dopant fluctuations (RDF), where the tails of the distribution scale with $W^{-0.5}$ $(L-\Delta L)^{-0.5}$ [17].

Comparing the overall RTN magnitude in nMOS and pMOS devices, pMOS devices tend to exhibit considerably higher RTN, which can be traced to both a higher number of observed traps and a larger single-trap ΔV_{th}. This result is interesting, as in terms of the RDF effects on V_{th}, pMOS devices generally exhibit less variation than nMOS devices [18]. Therefore, it would be expected that as transistor dimensions scale with new technology nodes and the comparative effect of RTN grows, pMOS devices would be affected more dramatically by RTN than nMOS devices.

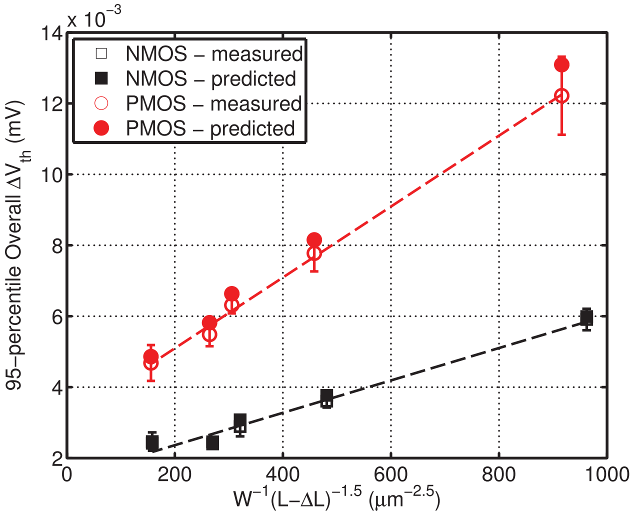

Fig. 10. Unfilled markers: 95th percentile measurements. Filled markers: predictions for the overall ΔV^*_{th} from nMOS and pMOS devices. Excellent agreement between measurement and prediction even in the tail of the distribution is demonstrated and a scaling trend inversely proportional to $W(L-\Delta L)^{1.5}$ is observed.

V. Conclusion

A comprehensive methodology for the characterization and analysis of RTN in a 45-nm bulk CMOS process was presented. This methodology was used to extract an empirical statistical model for the overall effect of RTN on the threshold voltage of small-area devices. Scaling of the individual RTN parameters was found to lead to a super-linear inverse scaling of the tails of the RTN distributions with respect to device area, where RTN amplitude was more sensitive to scaling of the effective device length. This conclusion, when combined with the observed skewness of the measured RTN distributions, indicated that RTN could become a dominant source of variability in future technology nodes.

Acknowledgment

The authors would like to thank Semiconductor Research Corporation for their support, which made this work possible.

References

[1] N. Tega, H. Miki, F. Pagette, D. Frank, A. Ray, M. Rooks, W. Haensch, and K. Torii, "Increasing threshold voltage variation due to random telegraph noise in FETs as gate lengths scale to 20 nm," in *Proc. Symp. VLSI Technol.*, Jun. 2009, pp. 50–51.

[2] J. Campbell, L. Yu, K. Cheung, J. Qin, J. Suehle, A. Oates, and K. Sheng, "Large random telegraph noise in sub-threshold operation of nano-scale nMOSFETs," in *Proc. IEEE Int. Conf. IC Design Technol.*, May 2009, pp. 17–20.

[3] T. Grasser, B. Kaczer, W. Goes, H. Reisinger, T. Aichinger, P. Hehenberger, P.-J. Wagner, F. Schanovsky, J. Franco, P. Roussel, and M. Nelhiebel, "Recent advances in understanding the bias temperature instability," in *Proc. IEEE Int. Electron Devices Meeting*, Dec. 2010, pp. 4.4.1–4.4.4.

[4] N. Tega, H. Miki, T. Osabe, A. Kotabe, K. Otsuga, H. Kurata, S. Kamohara, K. Tokami, Y. Ikeda, and R. Yamada, "Anomalously large threshold voltage fluctuation by complex random telegraph signal in floating gate Flash memory," in *Proc. IEEE Int. Electron Devices Meeting*, Dec. 2006, pp. 491–494.

[5] A. Ghetti, C. Compagnoni, F. Biancardi, A. Lacaita, S. Beltrami, L. Chiavarone, A. Spinelli, and A. Visconti, "Scaling trends for random telegraph noise in deca-nanometer Flash memories," in *Proc. IEEE Int. Electron Devices Meeting*, Dec. 2008, pp. 835–838.

[6] S. Realov and K. Shepard, "Random telegraph noise in 45-nm CMOS: Analysis using an on-chip test and measurement system," in *Proc. IEEE Int. Electron Devices Meeting*, Dec. 2010, pp. 28.2.1–28.2.4.

[7] K. Takeuchi, T. Nagumo, S. Yokogawa, K. Imai, and Y. Hayashi, "Single-charge-based modeling of transistor characteristics fluctuations based on statistical measurement of RTN amplitude," in *Proc. Symp. VLSI Technol.*, Jun. 2009, pp. 54–55.

[8] 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*, Dec. 2009, pp. 32.1.1–32.1.4.

[9] P. Fantini, A. Ghetti, A. Marinoni, G. Ghidini, A. Visconti, and A. Marmiroli, "Giant random telegraph signals in nanoscale floating-gate devices," *IEEE Electron Device Lett.*, vol. 28, no. 12, pp. 1114–1116, Dec. 2007.

[10] A. Ghetti, C. Compagnoni, A. Spinelli, and A. Visconti, "Comprehensive analysis of random telegraph noise instability and its scaling in deca-nanometer Flash memories," *IEEE Trans. Electron Devices*, vol. 56, no. 8, pp. 1746–1752, Aug. 2009.

[11] K. Sonoda, K. Ishikawa, T. Eimori, and O. Tsuchiya, "Discrete dopant effects on statistical variation of random telegraph signal magnitude," *IEEE Trans. Electron Devices*, vol. 54, no. 8, pp. 1918–1925, Aug. 2007.

[12] D. Fleury, A. Cros, K. Romanjek, D. Roy, F. Perrier, B. Dumont, H. Brut, and G. Ghibaudo, "Automatic extraction methodology for accurate measurements of effective channel length on 65-nm MOSFET technology and below," *IEEE Trans. Semicond. Manuf.*, vol. 21, no. 4, pp. 504–512, Nov. 2008.

[13] S. Realov and K. Shepard, "On-chip combined C-V/I-V transistor characterization system in 45-nm CMOS," in *Proc. Symp. VLSI Circuits*, 2011, pp. 218–219.

[14] H. Mueller and M. Schulz, "Individual interface traps at the Si–SiO$_2$ interface," *J. Mater. Sci., Mater. Electron.*, vol. 6, no. 2, pp. 65–74, 1995.

[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*, Dec. 2007, pp. 169–172.

[16] N. Tega, H. Miki, Z. Ren, and C. D'Emic, "Reduction of random telegraph noise in high-κ/metal-gate stacks for 22 nm generation FETs," in *Proc. IEEE Int. Electron Devices Meeting*, 2009, pp. 32.4.1–32.4.4.

[17] M. Pelgrom, A. Duinmaijer, and A. Welbers, "Matching properties of MOS transistors," *IEEE J. Solid-State Circuits*, vol. 24, no. 5, pp. 1433–1439, Oct. 1989.

[18] H. Takamizawa, Y. Shimizu, K. Inoue, T. Toyama, N. Okada, M. Kato, H. Uchida, F. Yano, A. Nishida, T. Mogami, and Y. Nagai, "Origin of characteristic variability in metal-oxide-semiconductor field-effect transistors revealed by three-dimensional atom imaging," *Appl. Phys. Lett.*, vol. 99, no. 13, pp. 133502-1–133502-3, 2011.

Simeon Realov (M'12) received the M.S. and Ph.D. degrees in electrical engineering from Columbia University, New York, NY, USA, in 2007 and 2012, respectively.

He is currently with Intel Corporation, Hillsboro, OR, USA.

Kenneth L. Shepard (M'91–S'M03–F'08) received the M.S. and Ph.D. degrees in electrical engineering from Stanford University, Stanford, CA, USA, in 1988 and 1992, respectively.

He has been with Columbia University, New York, NY, USA, since 1997.