A fast and test-proven methodology of assessing RTN/fluctuation on deeply scaled nano pMOSFETs

R. Gao[1], M. Mehedi[2], H. Chen[3], X. Wang[4], J.F. Zhang[2], X.L. Lin[1], Z.Y. He[1], Y.Q. Chen[1], D.Y. Lei[1], Y. Huang[1], Y.F. En[1], Z. Ji[5] (zhigangji@sjtu.edu.cn) and R. Wang[6]

[1]China Electronic Product Reliability and Environment Research Institute, Guangzhou, 510610, P. R. China [2]School of Engineering, Liverpool John Moores University, Liverpool, L3 3AF, UK [3]School of Microelectronics, Shanghai Jiaotong University, 200240, P. R. China [4]Department of Electronics & Electrical Engineering, Harbin Institute of Technology at Weihai, Shandong, 264209, P. R. China [5]National Key Laboratory of Science and Technology on Micro/Nano Fabrication, Shanghai Jiao -tong University, Shanghai, 200240, P. R. China. [6]Institute of Microelectronics, Peking University, Beijing, 100871, P. R. China

***Abstract*—Random Telegraph Noise (RTN)/fluctuation is one of the most serious reliability issues in modern deeply scaled CMOS. The current RTN characterization methods need to select devices and can only capture the fast traps, thus it is very difficult to predict and validate device long-term fluctuation behavior. A new fast and test-proven methodology of assessing RTN/fluctuation is proposed in this work. By using the Within Device Fluctuation (WDF), all the devices' fluctuation can be captured. Moreover, WDF can be well explained and simulated as a sum of all the As-grown Traps (AT) induced RTN.**

***Index Terms*--RTN, Long-term fluctuation, Within Device Fluctuation (WDF), As-grown Traps (AT).**

I. INTRODUCTION

Random Telegraph Noise (RTN) is regarded as one of the most serious reliability issues in scaled-down CMOS devices [1-9]. Valuable attempts have been made [10, 11], but several shortcomings remain to be tackled: (i) The time window (tw), used in a typical RTN test is short (e.g. ⩽1sec [3-5]) and cannot capture slow traps. Although RTN can be observed (Fig. 1a) within a short *time window* (tw), increasing tw gives a rising fluctuation which must be properly considered (Fig. 1b); (ii) Some early models rely on extracting capture/emission time of individual traps [2-4]. This cannot be carried out for a large percentage of devices, since the signal appears as a complex fluctuation, rather than clear and analyzable RTN [5]. The requirement for such device selection [7, 8] introduces uncertainty in the device-to-device variation (DDV) assessment; (iii) Although many methodologies for modelling RTN have been proposed, their predictive capability is difficult to validate.

In this work, a new methodology of assessing RTN/fluctuation on deeply scaled nano-scale pMOSFETs is proposed. The new method solves/alleviates all the 3 shortages mentioned above. (i) It measures fluctuation on all devices instead of selected devices with clear RTN. (ii) the testing time for each device is about 10 min. Moreover (iii) the long-term fluctuation under various Vg levels is well predicted. The improvement is ascribed to the handy definition of Within Device Fluctuation (WDF) and in-depth understanding of the different types of traps.

The measurement and definition of WDF is illustrated in Fig. 1b. A constant gate voltage (Vg) is applied on the DUT, the drain is biased at -100mV, drain current (Id) is continuously monitored at a 1 MSa/s sampling rate. In contrast to RTN analysis which deals with the raw Id data (Fig. 1a), WDF is defined as the subtraction of the "Upper Envelope" and "Lower Envelope" of the Id fluctuation, as shown in Fig. 1b.

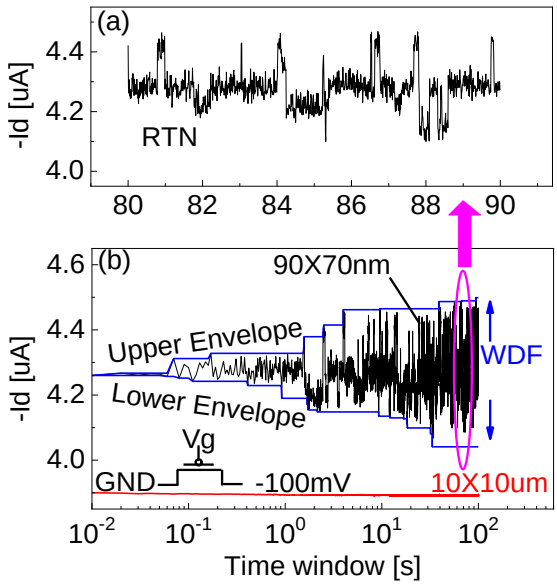

Figure 1. (a) RTN and (b) fluctuation with short and long time window respectively. WDF is defined as the subtraction of the "Upper Envelope" and "Lower Envelope" of the fluctuation. System noise is also measured on a 10um*10um DUT as a comparison.

II. DEVICES AND EXPERIMENTS

To validate the proposed methodology, planar pFETs of metal gate and HfO_2/SiON stack, fabricated at two different technology nodes are used in this work. Process A1 was fabricated at 45nm technology node with an Effective Oxide Thickness (EoT) of 1.45nm, while Process A2 was fabricated

This work was supported by *the Key Area R&D Program of Guangdong Province under Grant 2019B010145001* and *the National Natural Science Foundation of China under the grant no. 61927901*.

at 22nm technology node with EoT=1nm. Different treatments in respective technology nodes introduce different trap profiles in the dielectric of the DUTs. These profiles were utilized to identify the source of RTN/fluctuation in this work. Unless specified, A1 is used for illustration.

We have recently developed a test-proven As-grown Generation (AG) model which well predicts the aging kinetics and lifetime of nano-scale MOSFETs [12, 13]. The good prediction capability is ascribed to the in-depth understanding and subtle experimental separation of different types of defects. 3 types of defects are experimentally identified in the A-G model, as depicted in Fig. 2. Previous work [6] reveals WDF originates from the charging/discharging of defects in the gate dielectric. The vast majority of Generated Defects (GD) are very difficult to discharge once they are generated, even under a positive discharging voltage [14]. GD can thus be ruled out from the potential sources of WDF. The remaining two types of defects pre-exist on a fresh DUT and are both able to charge/discharge under stress/recovery phases. As-grown Traps (AT) capture holes without changing their energy levels; while Energy Alternating Defects (EAD), after capturing holes, shift their energy from ground level to a charged level, as illustrated in Fig. 2. We will show that AT is responsible for WDF from both theoretical and experimental side in section III.

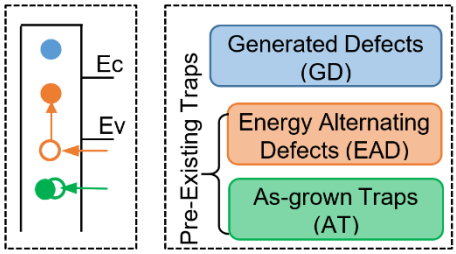

Figure 2. Illustration of the 3 types of defects in As-grown Generation (AG) model.

Since Generated Defects (GD) are irrelevant to WDF, its impact on measuring the charging kinetics of pre-existing traps needs to be removed. A test pattern as shown in Fig. 3 is designed for this purpose. To suppress the GD generation during the measurement of the pre-existing traps charging kinetics, each nano-scale device is firstly heavily stressed under a much higher voltage (compared to the highest charging voltage (Vgch)). Since GD follows a power law against both stress voltage and stress time [14], a stress under a high voltage for a certain period is equivalent to a stress under a lower voltage for a much longer stress time [15]. On a heavily stressed DUT, further generation during the Vgch measurements thus is prohibited. As long as we apply a positive Vgdisch to release all the pre-existing defects charged during the previous charging kinetics measurement, the DUT restores to its original state before charging and can be re-used to measure the next Vgch charging kinetics. This makes it feasible to measure charging kinetics under various Vgch on a single DUT, as shown in Fig.3. Charging kinetics is measured by intermittent *Id-Vg* curves from a three microseconds pulse edge at a logarithmic-incremental time sequence from 1 microsecond to 100 seconds, Vth is sensed at a constant current = 500nA*W/L. WDF at Vgch is measured right after Vgch charging kinetics without stopping Vgch stress. Since the DUT is already stabilized after 100 seconds of charging, influence from aging is excluded and "pure" WDF is captured.

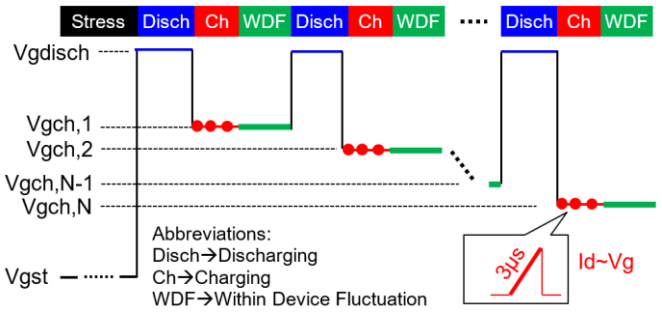

Figure 3. Test waveform to measure the Pre-Existing traps charging kinetics and WDF on a single DUT.

III. WITHIN DEVICE FLUCTUATION (WDF) SIMULATION AND TEST VALIDATION

A. *Capturing the right type of defects responsible for WDF*

Previous works [13, 16] reveal that there exist two types of traps with different charging mechanisms. By applying the procedure proposed in ref. 13, they can be reliably separated, based on their different charging kinetics (Fig. 4b). For nano-scale devices, these are the average of multiple devices (Fig. 4a).

In principle, WDF should be dominated by traps near the fermi level Ef in both neutral and charged states. These traps can readily charge and discharge under a constant Vg. Obviously AT meets this requirement as its energy level does not change after capturing a hole; while EAD is difficult to discharge after capturing a hole since its energy shifts to a lower energy level (Fig. 2) and should not contribute much to WDF.

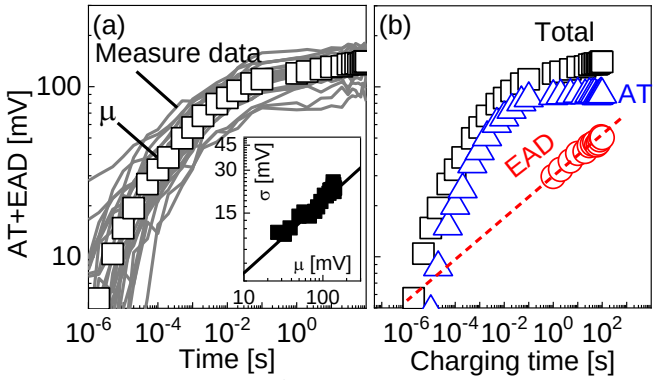

Figure 4. (a) Charging kinetics from multiple nano-scale devices (lines) and their averaged effect ('□'). Relationship between μ and σ is shown in the inset, where the average impact of a single trap η can be obtained. (b) Separation between AT and EAD following procedure in ref. 13.

To further confirm the source of WDF, a comparison of the averaged value of AT, EAD and WDF on process A1 and A2 is given in Fig. 5. Under the stress of gate dielectric electric filed (Eox) = 9 MV/cm, A1 exhibits the same amount of EAD but much higher AT compared to A2. WDF on A1 is

also much higher than A2, supporting the theory that WDF mainly originates from AT and is hardly affected by EAD.

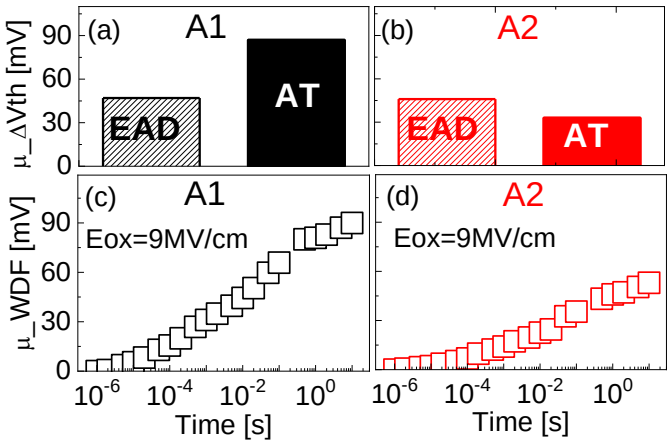

Figure 5. The comparison between EAD and AT for two processes (a) A1 and (b) A2 under an oxide electric field (Eox) of 9MV/cm. The corresponding averaged value of WDF from multiple devices are shown in (c) and (d) respectively.

B. WDF model construction

Since WDF and AT are from the same origin, WDF can be modeled from AT charging kinetics, which is a convolution of all the AT induced RTN. It is widely accepted that the statistical dynamics of an individual trap can be modelled with the non-radiative multi-phonon theory [17, 18] with Eqs.1&2:

$$\tau_c = 1/(n_s v \sigma_0 \cdot \exp(-\Delta E_B/kT)) \quad (1)$$

$$\tau_e = \tau_c \cdot \exp(\ (E_f - E_t)/kT) \quad (2)$$

where σ_0 is the average capture cross section, n_s is the carrier density in the inversion layer, v is the average velocity of the carriers, ΔE_B is the thermal activation barrier to capture a carrier, Ef is the fermi level and Et is the energy level of AT.

Based on Eqs.1&2, under a given Vg and fermi level Ef, AT far below Ef gives rise to RTN whose capture time is extremely short and much smaller than emission time. Their sum eventually forms the aging behavior as illustrated in Fig. 4a (grey lines). AT far above Ef gives rise to RTN whose capture time is extremely long and much larger than emission time, these defects will not be observed in short periods but will start to contribute as the time window evolves. These defects cause the logarithmic increase of WDF against time window, as shown in Fig. 1b and Fig. 5c&d. AT close to Ef gives rise to RTN which has comparable capture and emission time. These defects are the dominate source of the WDF we measured. By simulating the charging kinetics of AT from its energy profile and induced RTN, a complete image of WDF and its physics can be achieved.

Together with the defect-centric paradigm, AT charging kinetics on multiple nano-scale DUTs can be mimicked through monte carlo simulation, as shown in the flow chart of Fig. 6.

According to the defect-centric paradigm, on each DUT, Vth degradation, δV_{th}, caused by a single AT follows an exponential distribution. With the known average value of δV_{th}, η and δV_{th} of every AT can be generated with Eq. 3:

$$p(\delta V_{th}) = \eta^{-1}\exp(\ -\delta V_{th}/\eta) \quad (3)$$

where η is obtained from Eq. 4:

$$\eta = \sigma_{AT}^2/(2\mu_{AT}) \quad (4)$$

as shown in the inset of Fig. 4. The total number of AT, Nt, follows a Poisson distribution and can be generate with Eq. 5:

$$p(N_t = n) = N_{t0}^{n} \cdot \exp(-N_{t0})/n! \quad (5)$$

where N_{t0} is the average number of all the AT on a single nano-scale DUT. Although previous works [12, 13] reveal AT against the overdrive voltage (Vgov, Vgov=Vg-Vth) follows an exponential law instead of any cumulative distribution function (CDF), AT far away from Ef has little influence on WDF, a gaussian CDF which best fits our experimentally extracted AT profile (Fig. 7b) in the energy range of interest thus can be employed to get N_{t0}. The CVC simulator [19] is used to convert Vgov into trap energy Et-Ev. Once N_{t0} is known, μ_{E_t} and σ_{E_t} are also determined from the fitting. The energy level Et of each AT then can be generated with Eq. 6. For ΔE_B we follow early works [17, 20] and assume it follows a uniform distribution from 0 to ΔE_B_U. ΔE_B of each AT can then be generated with Eq. 7.

$$p(E_t) = \frac{1}{\sqrt{2\pi}\cdot\sigma_{E_t}} \cdot \exp\left(-\frac{(E_t - \mu_{E_t})^2}{2\sigma_{E_t}^2}\right) \quad (6)$$

$$p(\Delta E_B) = 1/\Delta E_B\_U \quad (7)$$

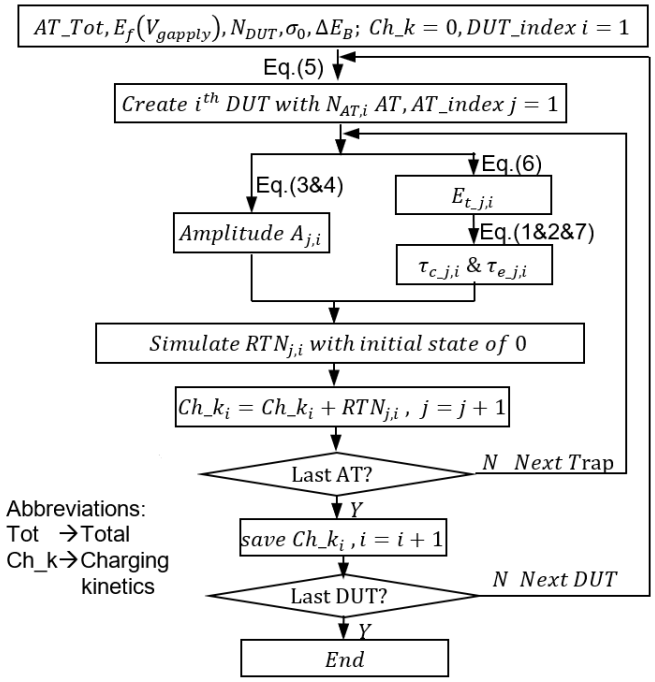

Figure 6. Flow chart of simulating AT charging kinetic based on the sum of all AT induced RTN.

C. Model parameter extraction

With the monte carlo simulation in Fig. 6, AT charging kinetics under an arbitrary Vg on each nano-scale DUT can

be obtained. Note the simulation only has two parameters to fit, σ_o and ΔE_B_U. All the other parameters are obtained from experiment data. By optimizing the σ_0 and ΔE_B_U values to get a best agreement between the simulated μ_AT charging kinetics and experiment data, σ_o and ΔE_B_U can be determined, as shown in Fig. 7a.

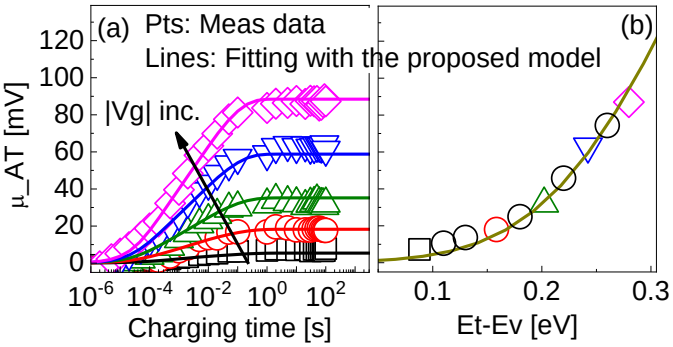

Figure 7. (a) The charging kinetics averaged from multiple-device measurements under different overdrive voltages, Vgov. (c) The measured energy profile of AT (pts) and the fitted curve with Gaussian distribution.

Fig. 8 shows the typical extracted values of η, σ_o and ΔE_B_U on different sizes of DUTs of both process A1 & A2. η shows the expected 1/(L*W) dependence (Fig. 8a), while both σ_o and ΔE_B_U are insensitive to DUT size (Fig. 10b&d). Remarkably, they are not sensitive to processes, either (Fig.10c&e), although the absolute value of WDF and AT is process-sensitive. It is speculated that different processes impact the trap density, but not trap properties. It should be noted that the extracted capture cross section $\sigma_o \sim 10^{-14}cm^2$ also agrees with the values obtained from previous independent tests in which hot holes were accelerated in the substrate to fill the trap directly [21].

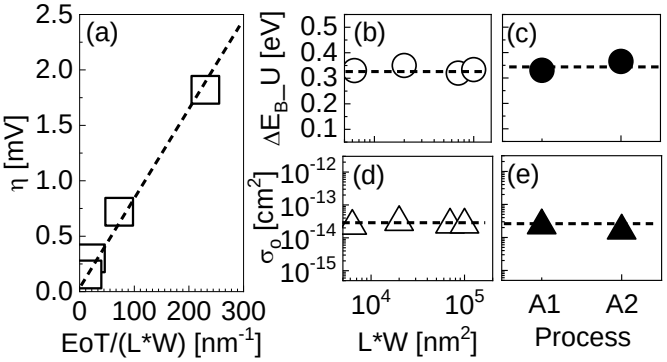

Figure 8. (a) Averaged single trap impact, η, increases with device area in the Takeuchi plot. (b) The upper boundary of the energy barrier, ΔE_B_U and (d) the capture cross-section, σ_0, shows insensitivity to the device area. Both ΔE_B_U and σ_0 also show little changes across different HKMG processes (c&e).

D. Model validation

The monte carlo simulation can mimic the test pattern in Fig. 3. WDF results can be simulated by picking the simulated AT charging kinetics beyond 100 seconds and then employing the "Upper Envelope" and "Lower Envelope" analysis. To validate the model, WDFs under |Vgov|=0.7V are measured on multiple devices and compared with 100 simulation results. Excellent agreement is achieved. The simulated WDF, averaged over multiple devices, under various Vg agrees well with test data for both processes, as shown in Fig.9 a&b. The relationship between the mean and the standard deviation from simulation also agrees well with test data, as shown in Fig. 9c&d. The comparison of the raw data in Fig. 9a under |Vgov|=0.7V is also given in Fig. 10 and a good agreement is also observed. It should be emphasized that the agreement in Fig. 9&10 is not a result of fitting. The type of tests for extracting model parameters is independent of and different from the WDF test used for validation. Since the model parameter extraction does not rely on the RTN analysis, the test time is shortened to ~10 mins per device, making it suitable for rapid assessment during process development.

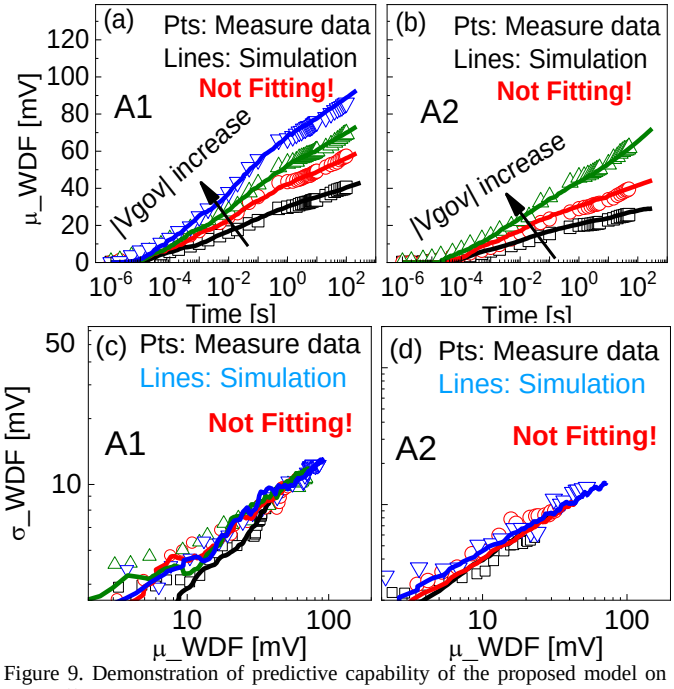

Figure 9. Demonstration of predictive capability of the proposed model on two different HKMG processes (A1 & A2). (a&b) the mean and (c&d) deviation of WDF from multiple-device measurements are compared with the prediction from 500 Monto Carlo simulation under the same condition. Devices from both processes are 90nm*70nm.

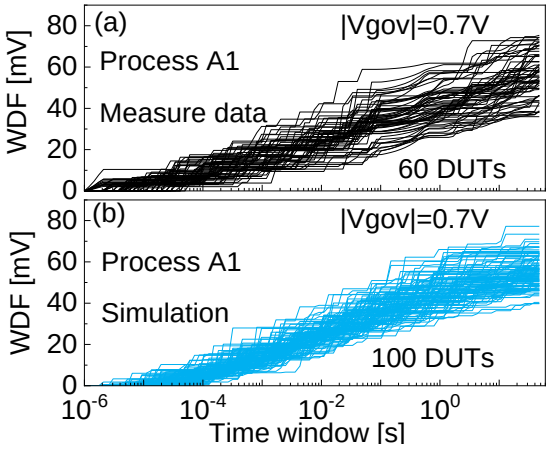

Figure 10. The comparison of WDF between (a) the experiment and (b) the simulation from the proposed simulation.

To compare simulation with tests further, Vg close to the threshold voltage is applied to 150 devices. The measured

signals can be divided into three groups: without analyzable RTN, with 1-trap RTN and with 2-trap RTN [22], as depicted in Fig.11a-c. The similar groups of signals can also be found in the simulation, as shown in Fig.11d&f under the same conditions as tests. Moreover, all devices were used to calculate the percentage of each group and simulation agrees well with tests, as illustrated in Fig.11g, further supporting the proposed methodology.

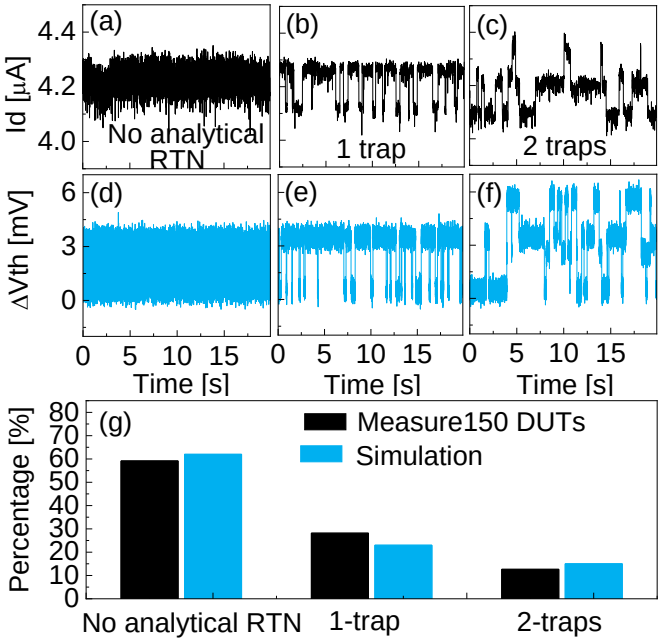

Figure 11. Different categories of RTN signals (no analytical, 1-trap and 2-trap) can be found in both (a-c) the measurements and (d-f) the simulations under the voltage of Vth0+0.2. System noise is taken into account in the simulation. (g) The comparison of the percentage of different RTN category occurrence between experiments and simulation.

IV. Conclusion

For the first time, this work proposes a new methodology for characterizing and modelling RTN/fluctuation and verifies its prediction capability by comparing the simulation with the independently measured data. The method is applicable to all devices and the testing time for each device is shortened to ~10 min/device, making it attractive for rapid process screening during development.

V. References

[1] A. Asenov, "Simulation of Statistical Variability in Nano MOSFETs," in *VLSI Symp. Tech. Dig.*, pp. 86-87, 2007.

[2] S. S. Chung, "The process and stress-induced variability issues of trigate CMOS devices," in *Proc. EDSSC*, pp. 1-2, 2013.

[3] K. Ota, M. Saitoh, C. Tanaka, D. Matsushita, and T. Numata, "Systematic study of RTN in nanowire transistor and enhanced RTN by hot carrier injection and negative bias temperature instability," in *VLSI Symp. Tech. Dig.*, pp. 1-2, 2014.

[4] H. Miki, M. Yamaoka, D. J. Frank, K. Cheng, D. Park, E. Leobandung, and K. Torii, "Voltage and temperature dependence of random telegraph noise in highly scaled HKMG ETSOI nFETs and its impact on logic delay uncertainty," in *VLSI Symp. Tech. Dig.*, pp. 137-138, 2012.

[5] C. Liu, L. Kyong Taek, L. Hyunwoo, K. Yoohwan, P. Sangwoo, and P. Jongwoo, "New observations on the random telegraph noise induced Vth variation in nano-scale MOSFETs," in *Proc. IRPS*, pp. XT.17.1-XT.17.5, 2014.

[6] M. Duan, J. F. Zhang, Z. Ji, W. Zhang, B. Kaczer, T. Schram, R. Ritzenthaler, A. Thean, G. Groeseneken, and A. Asenov, "Time-dependent variation: A new defect-based prediction methodology," in *VLSI Symp. Tech. Dig.*, pp. 1-2, 2014.

[7] 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 *IEDM Tech. Dig.*, pp. 28.3.1-28.3.4, 2010.

[8] A. Teramoto, T. Fujisawa, K. Abe, S. Sugawa, and T. Ohmi, "Statistical evaluation for trap energy level of RTS characteristics," in *VLSI Symp. Tech. Dig.*, pp. 99-100, 2010.

[9] Z. Ji, H. Chen, and X. Li, "Design for reliability with the advanced integrated circuit (IC) technology: challenges and opportunities," *Science China Information Sciences*, vol. 62, no. 12, p. 226401, 2019.

[10] M. Yamaoka, H. Miki, A. Bansal, S. Wu, D. J. Frank, E. Leobandung, and K. Torii, "Evaluation methodology for random telegraph noise effects in SRAM arrays," in *IEDM Tech. Dig.*, pp. 32.2.1-32.2.4, 2011.

[11] T. Tsuchiya, N. Tamura, A. Sakakidani, K. Sonoda, M. Kamei, S. Yamakawa, and S. Kuwabara, "Characterization of oxide traps participating in random telegraph noise using charging history effects in nano-scaled MOSFETs," *ECS Trans.*, vol. 58, no. 9, pp. 265-279, 2013.

[12] Z. Ji, S. F. W. M. Hatta, J. F. Zhang, J. G. Ma, W. Zhang, N. Soin, B. Kaczer, S. D. Gendt, and G. Groeseneken, "Negative bias temperature instability lifetime prediction: Problems and solutions," in *IEDM Tech. Dig.*, pp. 15.6.1-15.6.4, 2013.

[13] R. Gao, Z. Ji, S. M. Hatta, J. F. Zhang, J. Franco, B. Kaczer, W. Zhang, M. Duan, S. D. Gendt, D. Linten, G. Groeseneken, J. Bi, and M. Liu, "Predictive As-grown-Generation (A-G) model for BTI-induced device/circuit level variations in nanoscale technology nodes," in *IEDM Tech. Dig.*, pp. 31.4.1-31.4.4, 2016.

[14] R. Gao, A. B. Manut, Z. Ji, J. Ma, M. Duan, J. F. Zhang, J. Franco, S. W. M. Hatta, W. D. Zhang, B. Kaczer, D. Vigar, D. Linten, and G. Groeseneken, "Reliable Time Exponents for Long Term Prediction of Negative Bias Temperature Instability by Extrapolation," *IEEE Trans. Electron Devices*, vol. 64, no. 4, pp. 1467-1473, 2017.

[15] R. Gao, Z. Ji, A. B. Manut, J. F. Zhang, J. Franco, S. W. M. Hatta, W. D. Zhang, B. Kaczer, D. Linten, and G. Groeseneken, "NBTI-Generated Defects in Nanoscaled Devices: Fast Characterization Methodology and Modeling," *IEEE Trans. Electron Devices*, vol. 64, no. 10, pp. 4011-4017, 2017.

[16] P. Ren, R. Gao, Z. Ji, H. Arimura, J. F. Zhang, R. Wang, M. Duan, W. Zhang, J. Franco, S. Sioncke, D. Cott, J. Mitard, L. Witters, H. Mertens, B. Kaczer, A. Mocuta, N. Collaert, D. Linten, R. Huang, A. V. Thean, and G. Groeseneken, "Understanding charge traps for optimizing Si-passivated Ge nMOSFETs," in *VLSI Symp. Tech. Dig.*, pp. 1-2, 2016.

[17] 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," *Advances in Physics*, vol. 38, no. 4, pp. 367-468, 1989.

[18] K. Takeuchi, T. Nagumo, K. Takeda, S. Asayama, S. Yokogawa, K. Imai, and Y. Hayashi, "Direct observation of RTN-induced SRAM failure by accelerated testing and its application to product reliability assessment," in *VLSI Symp. Tech. Dig.*, pp. 189-190, 2010.

[19] J. R. Hauser and K. Ahmed, "Characterization of ultra-thin oxides using electrical C-V and I-V measurements," *AIP Conference Proceedings*, vol. 449, no. 1, pp. 235-239, 1998.

[20] R. Wang, M. Luo, S. Guo, R. Huang, C. Liu, J. Zou, J. Wang, J. Wu, N. Xu, W. Wong, S. Yu, H. Wu, S. Lee, and Y. Wang, "A unified approach for trap-aware device/circuit co-design in nanoscale CMOS technology," in *IEDM Tech. Dig.*, pp. 33.5.1-33.5.4, 2013.

[21] J. F. Zhang, H. K. Sii, G. Groeseneken, and R. Degraeve, "Hole trapping and trap generation in the gate silicon dioxide," *IEEE Trans. Electron Devices*, vol. 48, no. 6, pp. 1127-1135, 2001.

[22] M. D. Giles, N. A. Radhakrishna, D. Becher, A. Kornfeld, K. Maurice, S. Mudanai, S. Natarajan, P. Newman, P. Packan, and T. Rakshit, "High sigma measurement of random threshold voltage variation in 14nm Logic FinFET technology," in *VLSI Symp. Tech. Dig.*, pp. T150-T151, 2015.