
Trigger-When-Charged: A technique for directly measuring RTN and BTI-induced threshold voltage fluctuation under use-Vdd

http://researchonline.ljmu.ac.uk/10019/

Citation (please note it is advisable to refer to the publisher's version if you intend to cite from this work)


LJMU has developed LJMU Research Online for users to access the research output of the University more effectively. Copyright © and Moral Rights for the papers on this site are retained by the individual authors and/or other copyright owners. Users may download and/or print one copy of any article(s) in LJMU Research Online to facilitate their private study or for non-commercial research. You may not engage in further distribution of the material or use it for any profit-making activities or any commercial gain.

The version presented here may differ from the published version or from the version of the record. Please see the repository URL above for details on accessing the published version and note that access may require a subscription.

For more information please contact researchonline@ljmu.ac.uk

http://researchonline.ljmu.ac.uk/
Trigger-When-Charged: A technique for directly measuring RTN and BTI-induced threshold voltage fluctuation under use-Vdd


Abstract — Low power circuits are important for many applications, such as IoT. Device variations and fluctuations are challenging their design. Random telegraph noise (RTN) is an important source of fluctuation. To verify a design by simulation, one needs assessing the impact of fluctuation in both driving current, ΔId, and threshold voltage, ΔVth. Many early works, however, only measured RTN-induced ΔId. ΔVth was not directly measured because of two difficulties: its average value is low and it is highly dynamic. Early works often estimated ΔVth from ΔId/gm(Vg=Vdd), where gm is trans-conductance, without giving its accuracy. The objective of this work is to develop a new Trigger-When-Charged (TWC) technique for directly measuring the RTN-induced ΔVth. By triggering the measurement only when a trap is charged, measurement accuracy is substantially improved. It is found that there is a poor correlation between ΔId/gm(Vg=Vdd) and the directly measured ΔVth(Vg=Vth). The former is twice of the latter on average. The origin for this difference is analyzed. For the first time, the TWC is applied to evaluate device-to-device variations of the directly measured RTN-induced ΔVth without selecting devices.

Index terms: Random telegraph noise (RTN), Fluctuations, Yield, Within-a-device-fluctuation, Jitters, Positive charges, NBTI.

I. INTRODUCTION

As CMOS nodes scale down, the fluctuations induced by random charge-discharge of traps scale up. Smaller devices have larger statistical spread because of fewer traps per device and the larger impact of a single charge on them [1,2]. The increased number of devices per chip also leads to larger statistical spread [1,2] and high data transmission rate requires tight control of fluctuations [3]. Fluctuations have become a major concern for circuit design and have attracted many attentions recently [4-20]. It has been reported that current fluctuation in some fresh devices can be over the typical device lifetime criterion of 10% [5].

Fluctuations are commonly observed as the random telegraph noise (RTN) in the drain current, ΔId, under a given gate bias, Vg, and early works [5-13] have focused on them. ΔId allows probing individual traps and an analysis of their mean capture and emission time dependence on Vg gives the trap energy and spatial locations [5, 6, 8, 10]. This has improved our understanding substantially. There are, however, little direct measurements of the RTN-induced fluctuation in threshold voltage, ΔVth. This is because its measurement is difficult: the charge-discharge of traps for RTN is highly dynamic and the average ΔVth is typically low. As a result, the RTN-induced ΔVth often was either not given [5,11] or estimated from dividing ΔId by trans-conductance, i.e. ΔVth=ΔId/gm(Vdd) [6-10]. The accuracy of the ΔVth evaluated in this way was not given in these works [6-10].

To model the impact of RTN on the margin of SRAM [15] and the timing error [14], one needs both ΔId and ΔVth. For example, RTN in the pass transistor 1 in Fig. 1a can reduce the driving current by ΔId and slow down the Vg rise of transistor 2 in reaching its threshold voltage, Vtho, by Δt(ΔId). RTN in the transistor 2 can increase its Vth by ΔVth and results in a further delay, Δt(ΔVth). There is a need to obtain both accurate ΔId and ΔVth, therefore.

![Fig. 1. A schematic illustration of the impact of ΔId and ΔVth on timing: (a) circuits and (b) waveform. Vout switches when Vg=Vth, which is delayed by a lower charging current, Δt(ΔId), supplied through the transistor 1 and a higher Vth=Vtho+ΔVth of the transistor 2.](image)

The objective of this work is to develop a new Trigger-When-Charged (TWC) technique for directly measuring the RTN-induced ΔVth. By ensuring that the measurement is taken when traps are charged, the accuracy is substantially improved. It is found that the ΔId/gm(Vdd) correlates poorly with the directly measured ΔVth and the former doubles the latter on
average. The discrepancy originates partly from the device-to-device variation (DDV) of relative local current density beneath a trap at $V_g=V_{th}$ [16-19] and partly from the charge-induced mobility degradation [20].

Some deeply scaled devices have analyzable RTN signals in terms of extracting mean capture/emission time [11], while others can have a complex within-a-device-fluctuation [12]. The latter was deselected in some early works [10,13,16,17], making the real DDV of fluctuation unobtainable. The TWC developed in this work is applicable to devices with or without analyzable RTN signals and it will be used to evaluate the DDV.

II. DEVICES AND MEASUREMENT TECHNIQUE

A. Devices

The MOSFETs used in this work were fabricated by a 28 nm commercial CMOS process with a use $V_{dd}$ of 0.9 V. They have a metal gate and a high-k dielectric stack with an equivalent oxide thickness of 1.2 nm. The channel width and length are 135 nm and 27 nm, respectively. For comparison purpose, large devices of 3×1 µm were also used, which has insignificant DDV. All tests were performed at 125 ºC.

B. TWC technique

Difficulties with standard measure-stress-measure methods: For ageing-induced $\Delta V_{th}$ under stresses such as negative bias temperature instability (NBTI) [21,22] and hot carriers [23,24], the degradation is commonly measured at preset time. This is acceptable, as the $V_g$-acceleration used in the stress generally leads to a large-enough $\Delta V_{th}$ that is measurable and deterministic at a preset time. There are, however, two difficulties in applying this method to deeply scaled devices under use-$V_{dd}$, where $\Delta V_{th}$ mainly exhibits as Random Telegraph Noise (RTN). First, there are only a few active traps and the average $\Delta V_{th}$ is typically low. Second, charge-discharge of these traps are highly dynamic: they are often neutral at the preset time for measurement, as shown by the red circle symbols in Fig. 2, and would be missed by the measurement.

One way to avoid these difficulties is selecting devices that only have one trap, which induces a high enough $\Delta V_{th}$ (e.g. 20 mV) and its emission time is long enough (e.g. >1 sec) for completing the measurement [16,17]. This has improved our understanding of the interaction between a trap and the current. Such devices, however, are rare (e.g. ~10% [16]) and the required device selection precludes obtaining real DDV. The present work develops a new technique that removes the device selection and is applicable to all devices, so that the real DDV can be extracted.

Test procedure of TWC technique: Fig. 2a gives the $V_g$ waveform. After recording the reference Id-$V_g$ on a fresh device, the test starts by a ‘stabilization’ period of 40 sec under $V_g=V_{dd}=-0.9$ V. If there are any traps at deep energy level in a device, they will be filled during this period [25]. $\Delta Id$ under $V_g=-0.9$ V is then monitored for a period, e.g. 100 sec, as marked by ‘Id monitor’ in Fig. 2a. A sampling rate of 1 M/sec was used [26]. The trapping-induced up-envelope (UE) of $\Delta Id$ is obtained.

To measure the trapping-induced $\Delta V_{th}$, one must ensure that the measurement was taken when the traps are charged. This is achieved by setting the trigger level of the oscilloscope and the pulse generator for $V_g$ just below the UE, as shown in Figs. 2a&b. Once triggered, the pulse Id-$V_g$ (p-IV) is recorded in 3 µs to minimize discharge [25,26].

Although a sampling rate of 1 M/sec can be used to monitor $\Delta Id$ under a fixed $V_g=-0.9$ V, it only gives 3 points in 3 µs and is too slow for the p-IV. To have sufficient number of points for p-IV, a higher rate of 100 M/sec is used. The p-IV was repeatedly measured for 50 times and their average is used to reduce the system noise to ~1 mV.

![Fig. 2. The TWC technique.](image-url)
corresponding inset. The real $\Delta V_{th}$, however, should be evaluated at $V_{th}=-0.45\,\text{V}$ in Fig. 4a.

$V_{th}$ is extracted by extrapolating from the maximum $g_m$ point evaluated from $V_g=V_{th}$ at the point ‘A’ in Fig. 4a. We now compare the $\Delta V_{th}$ evaluated at $V_{gsense}=V_{th}$ (‘A’ in Fig. 4a) with that at $V_{gsense}=V_{dd}$ (‘B’ in Fig. 4a).

In the following ‘p-IV’ phase of Fig. 2a, when $V_{out}$ is above the ‘trigger level’ in Fig. 3c, the oscilloscope triggers and simultaneously sends out a signal to trigger the pulse generator for $V_g$. Both the pulse applied to the gate and the corresponding $V_{out}$ are captured, as shown in Fig. 3c. Two channels are needed here: channel 3 is at a fine scale to ensure capturing the small $V_{out}$ fluctuation with good accuracy and channel 2 is switched to a coarse scale to capture the whole p-IV. As a comparison, Fig. 3d shows an example triggered at a preset time that missed the trapped charge.

The UE in Fig. 2a can be caused by either a single trap or multiple traps. In the latter case, the UE results from the combined charges of multiple traps. This removes the need for selecting devices of a single trap and makes the method applicable to all devices.

The differences of this work from the typical BTI tests are that the p-IVs are only triggered when traps being charged and $V_g$-acceleration is not used here.

III. RESULTS AND DISCUSSIONS

A. A comparison between $\Delta I_d/g_m(V_{dd})$ and $\Delta V_{th}(V_{th})$

As mentioned in the introduction, early works [6-10] often estimated $\Delta V_{th}$ by $\Delta I_d/g_m(V_{dd})$, where both $\Delta I_d$ and $g_m$ were obtained under $V_g=V_{dd}$. This is effectively measuring the shift of IV at $V_g=V_{dd}$, as marked by the point ‘B’ in Fig. 4a and the corresponding inset. The real $\Delta V_{th}$, however, should be evaluated from $V_g=V_{th}$ at the point ‘A’ in Fig. 4a. In this work, $V_{th}$ is extracted by extrapolating from the maximum $g_m$ point and $V_{th}=-0.45\,\text{V}$ in Fig. 4a. The shift in $V_{th}$, $\Delta V_{th}$, at a given sensing $V_g$ is evaluated from $\Delta I_d/g_m(V_{gsense})$. We now compare the $\Delta V_{th}$ evaluated at $V_{gsense}=V_{th}$ (‘A’ in Fig. 4a) with that at $V_{gsense}=V_{dd}$ (‘B’ in Fig. 4a).

Fig. 4b plots $\Delta V_{th}(V_{th})$ against $\Delta V_{th}(V_{dd})=\Delta I_d/g_m(V_{dd})$ measured on 63 devices. Both of them have a large DDV, but the correlation between them is poor. For similar $\Delta I_d/g_m(V_{dd})$, $\Delta V_{th}$ can spread from its minimum to its maximum approximately. As a result, errors are large if $\Delta I_d/g_m(V_{dd})$ is used as $\Delta V_{th}$, so that it is essential to measure $\Delta V_{th}$ directly at $V_g=V_{th}$. Although both of them have maximum close to the typical device lifetime definition of 30–50 mV, the average $\Delta I_d/g_m(V_{dd})$ doubles that of $\Delta V_{th}$, as shown by the two dashed lines in Fig. 4b. This is because many devices have $\Delta V_{th}(V_{th})$ close to zero, but $\Delta I_d/g_m(V_{dd})$ are above 10 mV. The origin of the differences between these two will be analyzed next.

B. Effects of sensing $V_g$ on $\Delta V_{th}$

In Fig. 4a, the sensing $V_g$ for $\Delta V_{th}$ is -0.9 V for the point B and $V_{th}=-0.45\,\text{V}$ for the point A. Since the whole $I_d$-$V_g$ was measured, one can also extract the “apparent $\Delta V_{th}$” at other sensing $V_g$ by using $\Delta I_d/g_m(V_{gsense})$. The “apparent $\Delta V_{th}$” here is referred to the $\Delta V_{th}$ evaluated in this way under $V_{gsense}=V_{th}$. Typical examples obtained from different devices are given in Figs. 5a-e.
connected to the “External trigger in” of the pulse generator. (b) The Vout fluctuation is captured by both channels 2 and 3, as they are physically connected. (c) A screen-shot of the TWC p-IV measurement waveform. Channel 3 keeps its fine scale for accurate triggering, while channel 2 is switched to a coarse scale to capture the whole “TWC” p-IV. (d) A screen-shot of the traditional p-IV measurement at a preset time, where the trapped charge is missed.

The dependence of the apparent ΔVth on the sensing Vg has strong DDV, agreeing with that observed for single traps [16,17]. On one hand, Fig. 5a corresponds to Fig. 4a, where ΔVth increases monotonically with |Vg| and ΔVth at |Vg|=0.9 V is 6 times of the real ΔVth(Vth). On the other hand, ΔVth can also reduce by almost half over the same voltage range, as shown in Fig. 5b. There are also cases where (i) ΔVth is almost a constant (Fig. 5c); (ii) ΔVth increases initially and then reduces (Fig. 5d); and (iii) ΔVth decreases initially and then increases (Fig. 5e).

![Fig. 4. (a) Early works estimated RTN-induced ΔVth from ΔId/gm at Vdd=0.9 V (Point ‘B’), rather than directly measuring it at Vg=Vth (Point ‘A’). The two insets are enlarged p-IV around the two points. The black p-IV is reference and the blue p-IV is the TWC p-IV. (b) The poor correlation between ΔId/gm at Vdd and ΔVth at Vg=Vth. Each point was taken from a different device. The dotted lines mark the mean values.](image)

It is known that channel current can have a narrow percolation path near Vth and the impact of a charged trap on a deeply scaled device depends on the relative local current density beneath the trap [16-19]. This can explain the device-specific dependence observed in Fig. 5. As schematically illustrated in Fig. 6, for the device in Fig. 5a, the trap is located far away from the current percolation path at Vth, so that it has little impact and ΔVth(Vth) is low. The many close-to-zero ΔVth(Vth) points in Fig. 4b indicates that this is often the case. As Vg increases, the current becomes more evenly spread and its relative density under this trap rises, leading to the increase of ΔVth with Vg. As there is current flowing beneath each trap at Vdd, there is no close-to-zero apparent ΔVth in Fig. 4b, when evaluated by ΔId/gm(Vdd).

For the device in Fig. 5b, however, the trapped charge is on top of the current percolation path at Vth, resulting in a large ΔVth at Vth. As Vg increases, the current path is widened, so that the impact of the same charge on the device reduces and the ΔVth decreases with |Vg| in Fig. 5b. Similarly, the relative current density under the trap in Fig. 5c changes little with Vg and ΔVth is insensitive to Vg. The dependence of relative current density under a trap on Vg may not be monotonic, which can explain the behavior in Figs. 5d&e. For instance, in Fig. 5d, it may increase initially and then decrease. Alternatively, when there are multiple traps, some can behave like Fig. 5a and some like Fig. 5b. A combination of them can give the complex dependence in Figs. 5d&e.

![Fig. 5. Examples of the device specific dependence of the apparent ΔVth on the sensing Vg, Vgsense. (a)-(e) were obtained from five different devices. The apparent ΔVth at a Vgsense was obtained from the shift of TWC p-IV from the reference at Vgsense. The ΔVth is normalized against its value at Vgsense=Vth. As the lowest |Vgsense| is close to Vth, the data starts from ~1 in all devices.](image)

![Fig. 6. A schematic illustration of different impacts of traps at different locations on a device at threshold condition. The current can follow a percolation path under Vg=Vth. The trap in green corresponds to the device in Fig. 5a: it is away from the critical current path, so that it only has a small effect on the device at Vth. The trap in red corresponds to the device in Fig. 5b: it is on top of the current critical path and has a large effect on the device at Vth. Although the deeply scaled device-specific dependence of ΔVth on sensing Vg can be explained by the interaction between the trap and the relative local current density beneath it, there is also a device independent ΔVth dependence on the sensing Vg. For a large 3x1 μm device where DDV is insignificant, Fig. 7a shows that ΔVth also increases with](image)
On one hand, a more evenly distributed Id at higher Vgs sense allows more traps making an effective impact. On the other hand, the charge induced Columbic scattering causes mobility degradation [27,28], which lead to ΔId(mobility). When the apparent ΔVth is evaluated from ΔId(measured)/gm, the ΔId(mobility) is treated as if it was caused by ΔVth. In other words, the apparent ΔVth= ΔId(measured)/gm includes the contribution from mobility degradation to ΔId. As the effect of mobility degradation increases with |Vgsense|, it contributes to the increase in the apparent ΔVth for higher |Vgsense|.

C. Statistics

As there is hardly any information on the statistical properties of the directly measured RTN-induced ΔVth, especially in terms of its dependence on Vgsense, we report the DDV of this dependence here. Each line in Fig. 7b represents one device and the first impression is that the apparent ΔVth broadly increases for higher |Vgsense|. Although the ΔVth for some devices can reduce for higher |Vgsense| as shown in Fig. 5b, it is rare for a trap to be above a localized percolation path. As a result, the average (symbols in Fig. 7b) increases monotonically for higher |Vgsense|, which is partly driven by the mobility degradation.

The standard deviation, σ, is plotted against Vgsense in Fig. 7c. It can be divided into two regions: as |Vgsense| increases, σ decreases first and then increases. The minimum point is around 0.65 V. To explore this further, the relative variation, σ/μ, is also plotted in Fig. 7c. When |Vgsense|>0.65 V, σ/μ only rises modestly, so that the higher σ is mainly caused by the higher μ, as shown by the symbols in Fig. 7b. Below 0.65 V, however, σ increases and μ decreases for lower |Vgsense|, resulting in a rising σ/μ. When |Vgsense| lowers towards |Vth|, the current path becomes increasingly localized, leading to higher statistical variations, even though the trapped charges remain the same.

The cumulative distribution probability of ΔVth is given in Fig. 8a and σ is plotted against μ in Fig.8b for Vgsense=Vth. The RTN of nMOSFETs is smaller than that of pMOSFETs. σ follows μ by a power law with an exponent of ~0.5, agreeing with the prediction of Defect-Centric model [2, 16, 29]. According to this model, the average ΔVth induced by a trap, η, is,

\[ \eta = \frac{\sigma^2}{2\mu} \]

Using the fitted line in Fig. 8b, η ~ 3.2 mV is obtained for pMOSFETs. This η is ~2q/Cox approximately, where q is one electron charge and Cox the gate oxide capacitance. This agrees well with the value reported for the recoverable component of NBTI of pFinFETs [16], although the test samples used here are planar pMOSFETs from a different supplier. The average number of traps, N, per device is,

\[ N = \frac{\mu}{\eta} \]

For pMOSFETs, a μ ~ 12 mV in Fig. 4b gives N ~ 4.
exponential distribution. (b) Standard deviation versus mean. Lines show that the data follow the prediction of Defect-Centric model well with a power exponent of 0.5. The different pairs of (µ, σ) are obtained by varying the time window of "Id monitor" from 10 µs to 100 sec in Fig. 2a.

For nMOSFETs, the corresponding values are µ ~ 6.5 mV, η ~ 1.1 mV, and N ~ 6. When compared with pMOSFETs, the lower RTN in nMOSFETs is caused by smaller η. Although there are more traps in nMOSFETs, they are in the high-k layer and further away from the conduction channel and induce a smaller ΔVth [30].

IV. CONCLUSIONS

The conventional method of ‘Measure-Stress-Measure’ at preset time is inapplicable for the RTN-induced ΔVth, since the trap can be neutral when pulse IVs are taken. Early works estimate the RTN-induced ΔVth by ∆Id/gm at Vg=Vdd and its accuracy is not known. In this paper, we propose a new TWC method for directly measuring the real ΔVth at Vg=Vth. By setting the trigger level close to the upper envelope of trapping-induced ∆Id, it ensures that the pulse IV is taken when traps are charged.

Results show that there is no unique relationship between ∆Id/gm at Vg=Vdd and the directly measured ΔVth and their correlation is poor. The device-specific dependence of the apparent ΔVth on the sensing Vg originates from the DDV of relative local current density under a trap at Vth. Moreover, on average, ∆Id/gm(Vdd) doubles ΔVth(Vth) and the charge-induced mobility degradation through Coulombic scattering plays a role.

The TWC is applicable to devices with or without analyzable RTN signals. For the first time, it is used for assessing the statistical properties of the directly measured RTN-induced ΔVth. For the same trapped charges, it is found that σ has a minimum around |Vgsense|=0.65 V. The increase in σ when |Vg| lowers toward |Vth| is explained by an increased localization of current path. The DDV follows the Defect-Centric model. For the 135×27 nm devices used in this work, the average ΔVth induced per trap is ~3.2 mV for pMOSFETs and ~1.1 mV for nMOSFETs.

REFERENCES


