Microelectronics Reliability 81 (2018) 244–251

Contents lists available at ScienceDirect

Microelectronics Reliability

journal homepage: www.elsevier.com/locate/microrel

Review paper

Controversial issues in negative bias temperature instability

James H. Stathis[a,*], Souvik Mahapatra[b], Tibor Grasser[c]

[a] IBM T.J. Watson Research Center, Yorktown Heights, NY 10598, USA
[b] Department of Electrical Engineering, IIT Bombay, Mumbai 400076, India
[c] Institute for Microelectronics, Technische Universität Wien, Vienna, Austria

A B S T R A C T

In spite of 50 years of history, there is still no consensus on the basic physics of Negative Bias Temperature Instability. Two competing models, Reaction-Diffusion and Defect-Centric, currently vie for dominance. The differences appear fundamental: one model holds that NBTI is a diffusion-limited process and the other holds that it is reaction-limited. Basic issues of disagreement are summarized and the main controversial aspects of each model are reviewed and contrasted.

1. Introduction

For > 50 years, NBTI (Negative Bias Temperature Instability) has been recognized as a fundamental reliability issue for metal–oxide–silicon (MOS) transistors. The last broad reviews of this topic were published in this journal over 10 years ago [1,2]. In the intervening years, the number of publications has surged with 5900 publications in the period 2006–2016 [3]. Fig. 1 shows the number of annual publications with the phrase "negative bias temperature instability" since 1990, illustrating a dramatic increase beginning around 2001.

In NBTI, positive charges build up in the MOS gate insulator due to the application of negative gate bias (V_g), exacerbated by temperature (T). Some of the positive charge may dissipate when V_g is reduced. NBTI became an important concern with the introduction of nitrogen in silicon oxynitride (SiON) gate dielectrics and continues to remain a crucial issue for high-k/metal-gate (HKMG) technology.

The new generation of NBTI specialists have significantly advanced the experimental and theoretical sophistication, yet no consensus has formed on the physical mechanism(s) governing the kinetics of NBTI during DC and AC stress and for recovery after stress, in large and small area devices. On the contrary, two disparate viewpoints have taken hold, typically referred to as the Reaction-Diffusion or "RD" model-based comprehensive framework [4] and the "Defect-Centric" model [5,6]. Simply put, one holds that NBTI is a diffusion-limited process and the other holds that it is reaction-limited. While oversimplified, this description illustrates the fundamental level of disagreement.

Historically, the RD model focused on the kinetics of interface trap (N_{it}) generation to explain NBTI degradation (the stress phase). In the early 2000s, researchers began to turn their attention to the relaxation phase, i.e., the recovery of the V_t shift when gate bias is removed, and the Defect-Centric model grew out of this emphasis. The RD model began as a continuum model while the Defect-Centric model was based on the discrete behavior of individual traps. This difference in approach has shaped much of the discussion, but ultimately both models aim to consistently explain both large-scale average behavior and microscopic behavior.

Broadly speaking, the *critiques* of these two models can be summarized thusly:

(1) While the RD model can now very successfully describe a large variety of observed data over a broad set of experimental conditions, the validity of the underlying physical interpretation is questioned [5] because the model parameters conflict with established literature on H in Si/SiO_2 systems;
(2) The Defect-Centric model has a strong basis in physics, supported by microscopic measurement of discrete defects, but until recently [7–9] it has paid scant attention to interface state generation, and the ability to comprehensively describe NBTI stress data over a broad set of process and stress conditions is questioned [10].

We give more details below and address these critiques. While both models have had good success in explaining experimental observations, neither model has achieved the sine qua non of theory by uniquely predicting an effect in advance of its observation, nor of achieving a consensus that the other model is inconsistent with experiment (in spite of several published attempts [5,10,11]).

This paper is not a comprehensive review of NBTI. In particular, we will not address material dependence in detail, such as different gate dielectrics or channel materials, nor statistical distributions, although such topics may eventually help to prove or disprove a model. Here, we

* Corresponding author.
E-mail address: stathis@us.ibm.com (J.H. Stathis).

https://doi.org/10.1016/j.microrel.2017.12.035
Received 24 November 2017; Received in revised form 17 December 2017; Accepted 19 December 2017

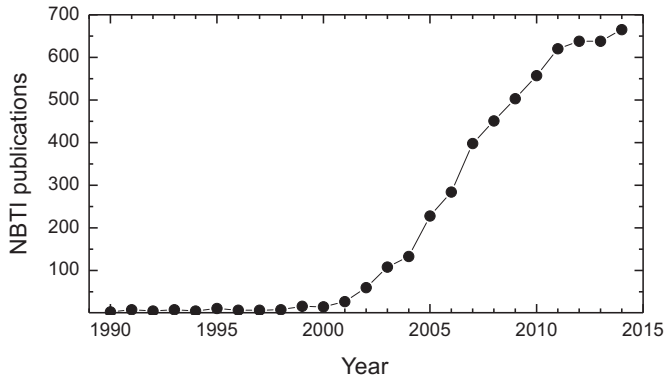

Fig. 1. Number of NBTI publications per year since 1990 [3].

outline the major controversies that continue to exist concerning the fundamental mechanisms.

2. Continuing controversies on basic issues

2.1. Roles of bulk traps and interface states

Experimental evidence shows that NBTI comprises both generated interface states (N_{it}) and hole trapping in bulk traps (N_{HT}, N_{OT}) [Reference [1] and references therein]. Interface states (N_{it}) are generally assumed to be responsible for changes in capacitance (CV), g_m degradation, and increased subthreshold swing as well as threshold voltage (V_t) shift when charged, and are frequently ascribed to the charge pumping (CP) or DCIV signals. Bulk traps can be either pre-existing hole traps (N_{HT}), or newly generated oxide traps (N_{OT}). They are often assumed to only affect V_t and give a rigid I_dV_g shift, although they can also contribute to DCIV and CP. However, there is still no consensus as to the most basic question of which component dominates the long-term V_t shift.

In the RD model, N_{it} generation – specifically, Si dangling bonds – dominates. Indeed, for state-of-the-art gate stacks, according to this model bulk trapping in pre-existing traps is insignificant, but new bulk traps can be generated under harsher stress conditions. The Defect-Centric model, on the other hand, focuses on hole capture and emission in bulk traps and border traps, which in this view dominate the process. Both models acknowledge that bulk traps can be generated under certain conditions, but the details of bulk trap generation are still being developed [7–9,12,13].

2.2. Interface state occupancy and recovery

During negative bias stress, interface states are not only created but they are positively charged; when the bias is removed, the occupancy changes and recovery (passivation) may happen [4,10]. The occupancy is important as only charged defects are sensed by measurement of V_t. The RD model includes an empirical interface state occupancy term with a stretched exponential spanning ~1μs to ~10s time constants, and passivation at longer times which is on the order of the stress time. Other researchers have claimed that neither component plays a significant role in normal recovery near zero bias [14]; occupancy changes of interface dangling bonds are too fast to be observed except by ultra-fast measurements, and passivation of interface states is either not observed at all, or only under certain conditions such as annealing at higher temperature.

2.3. Permanent vs recoverable components

Some of the early work on NBTI recovery decomposed NBTI into distinct recoverable (*R*) and permanent (*P*) parts [14]. This work claimed that *R* is due to hole de-trapping, and *P* is due to generated interface traps. To be more precise, it is now recognized that the "permanent" part is quasi-permanent on the time-scale of the experiments but can exhibit long-term recovery [15,16].

However, faster recovery has been reported in measurements like charge pumping, DCIV, and subthreshold slope, which are variously attributed to either N_{it} or oxide trap recovery [17,18]. In addition, the CP measurement itself accelerates recovery [16].

2.4. Measurement issues

Measuring NBTI turns out to be not straightforward, and no method exists which can measure NBTI directly and unambiguously without impacting the accumulated charges and defects and thus introducing some artifact. Therefore, most NBTI data leave some room for interpretation. In the following, the most commonly used measurement methods are discussed and their artifacts highlighted.

The conventional measure-stress-measurement (MSM) method first tries to establish a reference measurement, exposes the device to stress, and repeats the measurement to determine any changes. Typically, I_dV_g curves are measured to determine ΔV_t and Δg_m. However, as has been realized in the 2000s, I_dV_g measurements are inherently slow and the recovery occurring during that delay time leads to changes in the data.

In an attempt to minimize this delay, single-point measurements (sometimes called one point drop down) have been used which measure I_d only at a single gate voltage (typically near V_t) after a well-defined recovery/delay time t_{r0} rather than performing a complete I_dV_g sweep [19]. Conventional semiconductor parameter analyzers can perform this measurement in ~ms, and new instruments can perform it in ~10μs (so-called "*ultra-fast measurement*"). Different variants exist, e.g., measuring the shift in I_d at constant V_g, or the shift in V_g at constant I_d. In either case it is implicitly assumed that the major impact of NBTI is a rigid shift of the I_dV_g curve along the V_g axis and that the shape of the I_dV_g curve is only negligibly affected. This assumption may introduce some difficult-to-quantify errors for large degradation levels, although a final full I_dV_g can be used to estimate the impact.

In order to increase the amount of information, the recovery can be traced for a certain amount of time beyond t_{r0}, typically several decades, leading to what has been called the *extended MSM (eMSM)* method [20]. The impact of the measurement delay t_{r0} is clearly visible as the beginning of the recovery trace and only an extrapolation to $t_{r0} \rightarrow 0$ will give the true zero-delay degradation. This extrapolation requires some assumptions, in particular on the distribution of time constants. At the moment, it appears that it is impossible to measure fast enough to avoid this extrapolation as defects with time constants smaller than 1 μs exist which are difficult to measure.

In order to avoid this delay time problem, the *on-the-fly* (OTF) method was proposed which attempts to extract ΔV_t directly from the changes in the drain current I_d during stress [21]. Unfortunately, the on-the-fly method requires the measurement of a reference current $I_{d0} = I_d(t_{s0})$ at the start of stress, which has to be measured after a certain unavoidable minimum stress time t_{s0} which introduces an additional artifact [22–24]. In particular, t_{s0} has a strong impact on the power-law time exponent. Another drawback of the on-the-fly method is that I_d at stress voltage is more susceptible to changes in the mobility, which are difficult to separate from changes in the threshold voltage. In comparison, measurements done in the subthreshold regime depend exponentially on ΔV_t but only linearly on Δg_m, so most researchers no longer use the OTF method. In addition, OTF is not useful for measuring recovery.

Time-Dependent Defect Spectroscopy (TDDS) is a variant of eMSM, where repeated stress/recovery cycles on nanoscale devices are analyzed to build a 2D histogram of recovery step heights and times [25,26]. These steps occur with different heights, at stochastically distributed times, and the 2D histogram of recovery step heights and times reveals clusters corresponding to individual defects. TDDS