Microelectronics Reliability xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Microelectronics Reliability

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

Plasma process induced damage detection by fast wafer level reliability monitoring for automotive applications

D. Beckmeier *, A. Martin

Corporate Reliability Department, Infineon Technologies AG, Neubiberg, Germany

ARTICLE INFO

Article history:
Received 11 July 2016
Accepted 12 July 2016
Available online xxxx

Keywords:
Plasma process induced damage
Device reliability
Charge trapping
MOS gate oxide
Hot carrier stress
fWLR fast wafer level reliability monitoring

ABSTRACT

Plasma process induced damage (PID) poses a device lifetime risk to all semiconductor products containing MOS gate dielectrics. This risk increases for smaller technology nodes. In this work we will present how to protect automotive products from PID. Products need to be made robust against PID by design checks with antenna rules determined in technology reliability qualifications. Additionally, damage that is invisible at zero hour, i.e. in parameter or product tests, needs to be detected by fast wafer level reliability (fWLR) monitoring on the fully produced wafer. The application and details of different stress types for charging cases are presented and discussed.

1. Introduction

Plasma process induced damage (PID) poses a device lifetime risk to all semiconductor products containing MOS gate dielectrics [1–3,5]. In case of plasma process charging of a metal line connected to a MOS gate electrode, e.g. from via etching, depending on generated potential and on charging rate a current can flow through the oxide. This current can damage the oxide. The damage can be in the form of: an oxide breakdown, an increase in oxide leakage current, a zero hour parameter shift caused by charged defect states or a device lifetime reduction due to neutral oxide defects. Neutral defects can exist at the end of wafer processing because of annealing or other high temperature steps after the process in which the PID charging occurred. Due to their neutrality, they are invisible at zero hour in measurements without stresses, i.e. in parameter testing or product tests. They can get recharged during product lifetime by normal operation. The additional MOS transistor parameter drift caused by them can lead to an early fail of the product. For automotive products it is vital to ensure that significant MOS parameter drifts over the targeted lifetime are detected and the affected wafer material is not delivered to the customer.

Fig. 1 presents a way to ensure product safety with respect to PID. For automotive products, all levels should be applied. For less safety relevant applications the three upmost levels can be reduced, mainly in respect to number of test devices and monitoring sampling rate. The product designs need to be protected by a set of design rules for single layer antennae [4] for any oxide thickness to ensure that areas of layers that can get charged during processing are not too large. These rules need to be determined and verified in a technology reliability qualification for every device type and layer. This ensures a general robustness of the product to PID. For automotive technologies, also process deviations which can cause a critical lifetime reduction need to be identified. As PID is process tool dependent, it can occur systematically on few wafers of a lot or it can be located only in some areas of a wafer. For automotive products with stringent reliability targets, also partially affected wafers need to be identified. A way to do this efficiently is fast wafer level reliability (fWLR) [5,6] monitoring. For PID as part of fWLR monitoring, an analytic stress – a highly accelerated constant current stress (CCS) – is applied on the oxides from the gate and the drift of the device parameters is determined.

For fWLR monitoring, every device type is placed with and without (reference device) a stacked antenna (see Fig. 2) in the exact size of the design rules for that device type in the wafer kerf (or a drop-in area) on productive wafers.

By applying a constant current stress and comparing parameters before as well as after stress of antenna and reference device, a charging event can be detected and the lot can be stopped. In the next sections will be shown that this event can be threatening the targeted product lifetime, but it does not have to be. This is why the effect on the lifetime needs to be determined by a more quantitative WLR stress (top level in Fig. 1) like Hot Carrier Stress (HCS) or Bias Temperature Stress (BTS) on the affected wafers. For thin oxides with a thickness in the direct tunneling regime, also a Time Dependent Dielectric Breakdown Stress can be necessary. Which stress type to use should be chosen depending on which stress the affected device has a smaller lifetime margin in, as determined in device reliability qualification. As wafer chucks can usually

* Corresponding author.
E-mail address: Daniel.beckmeier@infineon.com (D. Beckmeier).

http://dx.doi.org/10.1016/j.microrel.2016.07.129
0026-2714/

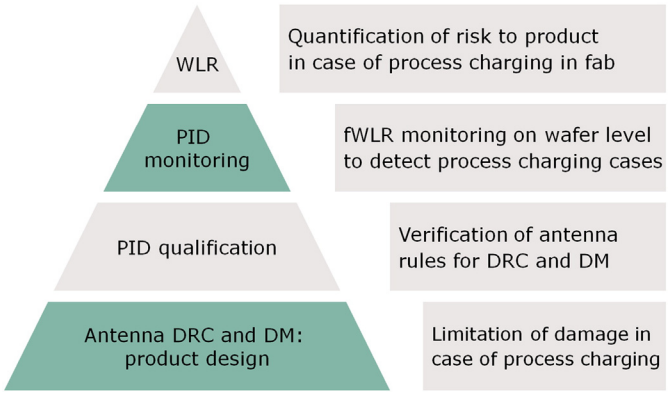

Fig. 1. Product safety with respect to PID is ensured by the levels in this pyramid. Each level depends on the level below it and means an increase in reliability and quality assurance.

not be heated in productive environments and kerf heater structures are usually not available for antenna devices [7], BTS on PMOS antenna devices is not feasible most of the times. Generally the resulting drifts from at least two stress conditions should be analyzed to verify the HCS or BTS lifetime models from the device reliability qualification. The wafers are scrapped in case the products have a lifetime risk. They are released if there is no risk. In both cases the tool causing the PID excursion needs to be identified and corrected as continuous improvement of production.

The effect of the stress level of the CCS during fWLR on detection sensitivity as well as more product relevant WLR HCS stresses will be presented and discussed in the following sections.

2. Analytic PID stresses in fWLR

To be able to measure large enough statistics, the fWLR CCS needs to be very fast and sensitive for the detection of damage. The effects of injected stress current density (J) and stress time need to be determined. The stress bias is applied to the MOS gate with respect to source, drain and well to inject a constant current density at room temperature as shown in Fig. 2. A stress algorithm controls the gate stress voltage in order to inject a constant current for a few seconds per stress step. After each step the transistor parameters are measured and then the next step is applied. Timings of this are kept strictly identical for all measurements to avoid variation of recorded MOS transistor parameters because of stress length and to avoid variable recovery effects of the parameters after the stress. As stress and measurements are performed at room temperature, and parameter measurements are rather quick (in total under 1 s per step), a recovery as known from BTS measurements should not change the relative drift behavior of the parameters [8]. For the following correlation measurements a large number of identically damaged devices need to be available. For this, technology reliability study wafers with MOS devices with stacked antennae sized far above product design rules were identified. Those devices showed an identical parameter drift after one fixed CCS step across the whole wafer and between wafers, respectively. P-MOSFETS with an oxide thickness of about 5 nm were chosen for the measurements, because this thickness range tends to be most sensitive to PID damage [2].

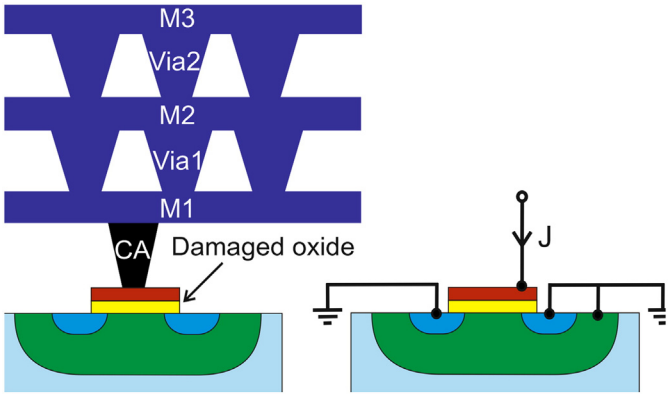

Fig. 2. Schematic of set of PID monitoring test structures consisting of a MOS transistor with a stacked antenna containing metal and via layers following the product design rules on the left and an identical device without antenna as reference structure on the right. The application of the constant current stress is depicted for the reference device.

Fig. 3 shows the drift of the linear threshold voltage (Vt) of PMOS transistors without antennae ("references") for four different J values versus the total injected charge on a logarithmic scale. Each point represents the median value for at least five devices. The injected charge is calculated from the injected J and the effective stress time per stress step. That means that the rise times of the stress voltages were taken into account. They can vary due to different capacitive charging currents into antenna and probing pad. It can be clearly seen in Fig. 3 that because of the CCS the Vt degrades with increased injected charge. The amount of degradation depends on the stress level. The higher the stress amount the larger the Vt degradation becomes.

The measurement data recorded with identical measurement conditions for MOS transistors with antennae from the same wafers are shown in Fig. 4.

In comparison to Fig. 3 the Vt drift values of the antenna devices are generally higher than for the references. The curves also demonstrate that the difference of Vt drift between reference and antenna increases for a fixed injected charge (e.g. 0,1 C/cm^2) with increasing stress current (e.g. from 160 mV at level 1 to 320 mV at level 4). The antenna Vt drift curves become steeper for a larger total amount of charge injected. This finding will be discussed in Section 4.

The forced voltage at the end of each stress step is presented in Fig. 5 for reference devices as line and for antenna device as symbols for the four stress levels.

An increase with larger current density is visible as expected for the tunneling mechanism. However, antenna and reference device show almost identical values, also for large total injected currents. This is very important for the PID monitoring as it means that the device oxides were stressed equally. Because of this the parameter drifts can be can be compared for PID detection.

It can be concluded from Figs. 3 and 4 that for monitoring purposes, a very short stress at a very high J yields a superior PID sensitivity

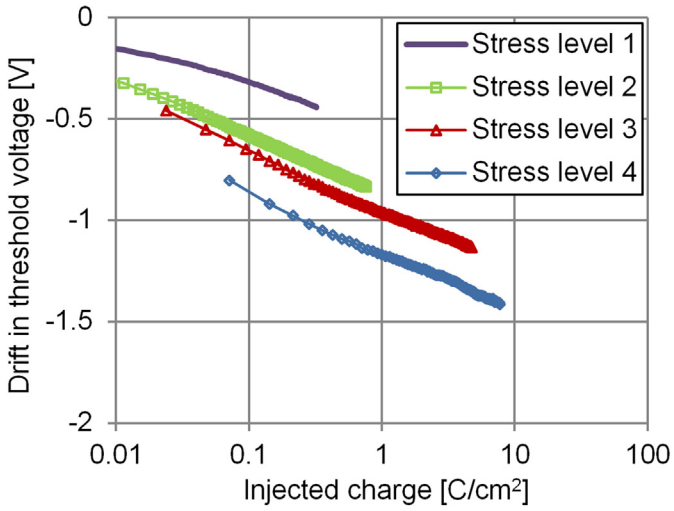

Fig. 3. Drift of the Vt with respect to its zero hour value versus the cumulative injected charge for identical pMOS reference devices for four stress current density levels 1–4 (see legend) applied on the gate in channel inversion.

Please cite this article as: D. Beckmeier, A. Martin, Plasma process induced damage detection by fast wafer level reliability monitoring for automotive applications, Microelectronics Reliability (2016), http://dx.doi.org/10.1016/j.microrel.2016.07.129