Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Microelectronics Reliability 44 (2004) 1209–1231

MICROELECTRONICS RELIABILITY

www.elsevier.com/locate/microrel

An introduction to fast wafer level reliability monitoring for integrated circuit mass production

Andreas Martin *, Rolf-Peter Vollertsen

Central Reliability Methodologies, Infineon Technologies AG, Otto-Hahn-Ring 6, 81739 Muenchen, Germany

Received 7 April 2004
Available online 5 June 2004

Abstract

The continuous verification of process reliability is essential to semiconductor manufacturing. The tool that accomplishes this task in the required short time is the fast wafer level reliability monitoring (fWLR). The basic approaches for this task are described in this introductory overview. It summarizes sampling plans, discusses the feasibility of using fWLR for screening and describes the data assessment and application of control cards. Beyond these general topics many of the fWLR stress methods are described in detail: Dielectric stressing by means of an exponential current ramp is compared to ramped voltage stress. Especially for thin oxides the methods differ regarding the soft breakdown detection and the time they consume. Another task of fWLR is the detection of plasma induced damage, which can be achieved by applying a revealing stress to MOSFETs with antenna. The design challenges of the structures and the test method as well as the data assessment are described in detail. An important section deals with fWLR for interconnects. In this section the appropriate test structures (including thermal simulations) are illustrated and fast electromigration stresses are discussed and the details of standard wafer level electromigration accelerated test (SWEAT) are included. For contacts and vias a simple method to check reliability is presented. Finally the monitoring of device reliability is treated. It is shown that using indirect parameters that correlate well to standard parameters such as the drain current can be beneficial for fWLR. For both, the interconnects and the devices, it is essential to have locally heated test structures in order to keep the stress time low. The detection and verification of mobile ions can also be performed with such a self-heated structure. For the described methods examples are given to illustrate the usefulness.

1. Introduction

Semiconductor manufacturers worldwide have introduced many different tools to control and monitor their product quality. Closely linked to a good quality is the yield of the manufacturing process. Only a process with high yield and good quality guarantees a profitable business. In the areas of design and layout of the circuit, the technology development or wafer processing and the packaging, various different tools are implemented for the monitoring and the control of the quality. Here in this paper the area of interest is the technology and wafer processing. Many control and monitoring tools which are directly applied before the wafers are diced are well known such as: process in-line monitors, process control monitoring (PCM), process qualification, reliability control monitor (RCM), quarterly monitoring (or re-qualification), wafer burn in and wafer product test. Most of these tools are implemented in the fabrication sites. Set up and the choice of methods depend on the fab situation and the stability of the process. Of these tools the RCM and the corresponding fast wafer level reliability (fWLR) monitoring is discussed in detail in this fWLR introduction.

Continuous monitoring and controlling of process reliability is an essential task during technology development, process ramp up and after process qualification

* Corresponding author. Tel.: +49-89-234-45257; fax: +49-89-234-9557386.
E-mail address: andreas.martin@infineon.com (A. Martin).

0026-2714/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.microrel.2004.04.008

during high volume production. It is assumed that during integrated circuit (IC) mass production the process reliability characteristics improve due to continuous learning and improvement of the process or at least do not change at all with respect to process qualification results. In reality process variations or misaligned process tools might cause reliability deviations. The instabilities which are responsible for a process reliability degradation can often not be identified by process in-line monitors or PCM. Therefore, an appropriate tool for this task is required. fWLR stressing, i.e. very short, highly accelerated stresses on RCM test structures integrated in the scribe line of product wafers, can be specifically employed to monitor the process reliability [1]. Regular fWLR monitoring is able to highlight reliability changes of process steps or tools and can indicate violations of the product reliability target when extrapolation models are available. In case of a reliability degradation fWLR is a trigger for root cause investigations and subsequent for process improvement, corrective actions or Burn-In. Also built in reliability (BIR) as a method for the continuous improvement of the process reliability is mainly based on a well established fWLR approach [2–4].

Originally the WLR monitoring idea and its implementation had been proposed in the literature by several people such as Crook [5], Turner [6], Messick [7] for various reliability problems and was put into integrated circuit mass production by many different semiconductor manufacturers [7–11]. The currently used fast WLR monitoring is a direct development of accelerated reliability testing on wafer level of some 10 years ago. In Fig. 1 the characteristics of accelerated reliability stressing is depicted and compared to the targeted product reliability. The time to failure is displayed as a function of the stress acceleration where the operating conditions can be found on the left side of the *x*-axis. The product itself must function under operating conditions for a specified lifetime. The typical lifetime (of more than $10\text{y} = 3.15 \times 10^8$ s) depends on the actual product operating conditions and thus varies to some degree with operation voltage and temperature as illustrated in Fig. 1. Product life tests and also long term reliability stresses are usually carried out under slightly accelerated stress conditions on package level with stress times from 10^4 to 10^8 s. Long term reliability stresses are used in process qualifications for extrapolation model parameter extraction and when new materials and/or process steps are introduced first time into a technology.

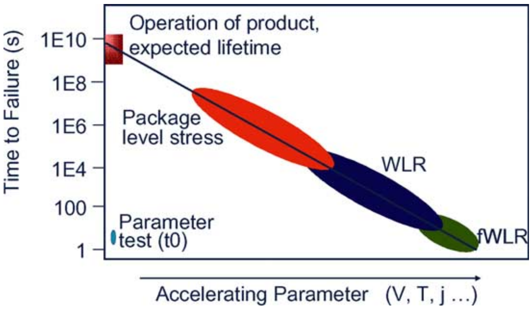

Fig. 1. Comparison of fWLR to other accelerated tests and targeted product lifetime.

For faster feedback to process development higher accelerated tests on wafer level are performed when the degradation mechanism is known to be the same as on package level, e.g. for metal line electromigration, gate oxide degradation and device reliability. WLR stress measurements are commonly applied during process qualification in addition to the package level tests. Stress times can vary from 10 to 10^5 s. Lifetime projections based solely on data of highly accelerated WLR stresses include some degree of uncertainty. However, if they are backed up by package level stresses they are very useful and save measurement time for lifetime extrapolations. In case of gate oxide reliability a lot of publications indicate that models at low electric field are also valid at electric fields of WLR stresses [12–14]. Also for electromigration measurements it has been shown that Black's Law [17] is valid for package level as well as fWLR monitoring [18].

Further increase of the acceleration on wafer level is possible for many mechanisms, however, the quantitative conclusion in some cases has to be replaced by a qualitative estimate. Stress times for fWLR can range from 1 to 100 s. In order to minimize the stress time special stress and measurement sequences are required in addition to a special test structure design [19]. At the end of a process qualification ideally fWLR measurements are performed on the qualification hardware. This result can serve as reference for sufficient reliability performance. fWLR data taken during high volume production can be compared to this reference with the help of SPC [2]. In this case no data extrapolation is required.

The short measurement time of fWLR monitoring allows a reasonably high sampling rate at the end-of-line test. The availability of the necessary RCM test structures in the scribe line of productive wafers provides also a fast feedback for process development and improvement if reliability problems occur. fWLR monitoring tests are also more economical in comparison to regular reliability tests on test chips which requires extra test wafer fabrication.

In comparison, usually, PCM tests measure the actual device parameters at time zero. Those parameter tests which are located in Fig. 1 in the lower left corner do not include any time dependence or degradation characteristics. Clearly PCM tests reflect the time zero device functionality characteristics but do usually not highlight reliability risks. Therefore, fWLR monitoring is considered to be an essential tool of the in-line test for reliability aspects in addition to PCM tests.

In comparison to quarterly reliability monitoring or quarterly re-qualification the regular weekly fWLR monitoring has the large advantage of giving a much faster feedback into production. Quarterly reliability monitoring or quarterly re-qualification can also identify reliability risks but in nearly all cases it is impossible to correct the root cause for it in time without having packaged products failing in a Burn-In or even field returns.

This paper summarizes the various fWLR methods that are successfully applied in mass production worldwide for addressing the most dominant technology reliability risks. Necessary test structures and critical design issues, test details, data screening and analysis and examples of successful problem detection will be presented and discussed. Some general considerations on fWLR monitoring such as sampling frequency are discussed first. The technical aspects are discussed in the following order: dielectric reliability, plasma induced damage, electromigration and device degradation.

2. Sampling frequency for fWLR monitoring

The testing frequency of fWLR monitoring is a key element for:

- the detection of processing problems,
- the correct interpretation of recorded data.

In general, the sampling of fWLR monitoring should correspond to the state of the process performance. For process development and/or the ramp up of the process a high sampling is recommended in order to get sufficient reliability information in a short time. However, for a stable and mature process the sampling frequency can be minimized, e.g. one complete stress cycle on one lot per week and technology. But for the following cases, listed below, which can occur during integrated circuit mass production the sampling must be increased according to an established out of control action plan (OCAP) which organizes adjustments of sampling and further actions. OCAP can be triggered by:

1. sudden maverick lot,
2. "continuous un-stable" process, no corrective action identified yet,
3. process or recipe change,
4. tool change, maintenance,
5. WLR-tester hardware/software up-date.

The temporary increase in the sampling can affect all reliability aspects (cases 1, 2 and 5) or the sample increase can be restricted to a subset of the fWLR monitoring tests (cases 1–4). The sampling frequency can be varied by the following three aspects:

- the sample size per wafer,
- the number of wafers per lot,
- the number of lots dependent on the total wafer starts per week.

Depending on the reliability mechanism the sampling size per wafer can vary typically from 3–5 to 20–25 samples. Surely, reliability risks with extrinsic (defect related, bimodal) failure modes require large sample sizes in order to gain an insight when and where on the wafer the extrinsic mode or the earlier mode appears. For wafer mapping of data in general a large sample size is needed. Typically, dielectric stress measurements are performed with a large sample size since the early extrinsic fails of the dielectric reliability are of interest and should be monitored. The dielectric stress measurements include the gate oxide reliability (also other dielectrics) as well as the plasma induced damage (PID) tests.

Table 1 summarizes the possible control modes, the fWLR frequency, the sampling on the wafer and the decision criteria for corrective actions. The control mode

Table 1
Definition of fWLR sample size, corrective actions and comparison of underlying reasons for the implementation of fWLR monitoring into a process

Control mode	fWLR frequency	Typical sampling per wafer	Decision criteria triggering corrective actions
Process control for reliability	1–5 Lots per week and technology dependent on wafer starts per week, 10–15 wafers per lot	15–25 Devices for dielectrics, 3–10 devices for other stresses	Statistically significant deviation (SPC criteria), root cause investigation—process improvement
Screening maverick lots	Every lot, 3–5 wafers per lot	3–15 Devices for dielectrics, 3–5 devices for other stresses	Reliability limit—re-measuring complete lot with higher sampling—scrap wafers or lot
Screening wafers	Every lot and every wafer	3–5 Devices for dielectrics, 1–3 devices for other stresses	Reliability limit—re-measuring wafers with higher sampling—scrap wafers

is the underlying reason for the implementation of fWLR monitoring. Three categories exist and are displayed in Table 1.

The first category "Process control for reliability" represents the standard mode for a fWLR monitoring sampling plan. It includes the testing of 10–15 wafers per lot selected for fWLR with the complete fWLR plan and where needed a large number of samples per wafer. In case of any violations the OCAP describes and triggers further actions. The aim is the detection of process instabilities, the immediate feedback into the process, the identification and the adjusting of the affected process tools. The main goal of fWLR monitoring is that the process keeps a stable reliability on a high level and as a result ensures a minimum of field fails from the customers.

The second category "Screening maverick lots" includes the fWLR testing of 3–5 wafers per lot. The underlying task of this control mode is the identification of a lot which shows process reliability problems. Since only a few wafers are regularly tested per lot an OCAP must trigger the testing of the remaining wafers of the lot and other corrective actions in case of a reliability target violation. In IC mass production this control mode will result in a cost intensive in-line measurement cycle when the complete fWLR plan is applied. However, when a single reliability risk is identified and the corresponding reliability stress sequence is time optimized then fWLR monitoring of each lot could still be economically even for mass production.

The third category "Screening wafers" requires that every single wafer of the production is monitored with a limited sample size per wafer. Clearly the aim of this control mode is to identify single wafers which do not meet the targeted process reliability and subsequently to down grade or scrap them according to the established OCAP. This is also a very cost and resource intensive measurement task. Additionally, it should be mentioned that some defect related reliability risks cannot be assessed on the basis of a few samples on one monitored wafer. In other words, the category, "Screening wafers", represents the least practical sampling strategy of a fWLR monitoring scheme.

3. fWLR data assessment

In any of the above sampling cases the fWLR data assessment over time should be performed using SPC-like charts with an upper (UCL) and lower control limt (LCL) [2]. For most reliability aspects only one control limit is critical and triggers corrective actions. An adjusted SPC chart and method is needed for fWLR. Note that also clear deviations to better reliability are worthwhile to be noticed and understood to reduce performance trade off. The SPC chart indicates any deviation from the usually expected process reliability as it is indicated in Fig. 2 while small deviations are tolerated as the usual process variation. In case of violations the established OCAP defines the subsequent action items and corrective measures. It is possible for some degradation mechanisms such as electromigration lifetime, hot carrier lifetime and gate oxide defect density that a reliability limit can be given which corresponds to the product target. This limit represents a measure for quantitative assessment beyond the LCL or UCL.

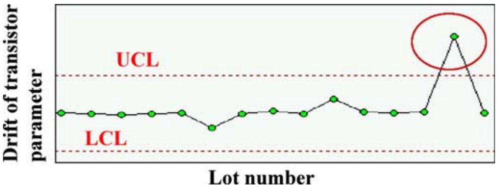

Fig. 2. Example of control card following SPC rules. A clear violation of the trend is marked with a circle when the lot data is above the upper control limit (UCL).

Care has to be taken when assessing the raw data in order to extract the data points for the control card. A step by step approach has proven useful.

1. The first step includes the data assessment of integrity tests before the actual fWLR stress and verification measurements after the stress was completed. This is a JEDEC conform methodology (e.g. in [31]) to avoid misleading data points from e.g. initially broken test structures or from malfunctioning stress sequences.
2. In the second step all necessary calculations are performed in order to get the reliability parameters that can indicate a reliability degradation in a cumulative plot.
3. The third step consists of the representation of the relevant fWLR data in a cumulative failure probability plot. From the distribution characteristic parameters are determined for each reliability mechanism. For dielectric reliability assessment the Weibull plot is applied and generally accepted in literature [15]. Electromigration results are usually displayed in log-normal plots [16]. Other parameters such as a drift of a transistor parameter or the change in leakage current is usually plotted as cumulative plots with a linear percentage scale but dependent on the parameter a log- or linear-scale for the drift value is used.
4. The fourth step is to extract the key parameters from the cumulative plot and represent this result in a control card as it is presented in Fig. 2.

This four-step approach will become clearer when the technical aspects of the fWLR measurements will be described in detail in the following sections.

4. Dielectric reliability

The test structures used for dielectric reliability investigations can have various shapes and geometries: large capacitors with rectangular plate, transistors, arrays of transistors or arrays of small unit cell capacitors [20], structured capacitors with e.g. fingers, serpentines. Output data can be strongly influenced by the edge to area ratio of a test structure. In general it is assumed that defects and intrinsic weaknesses are randomly distributed across the area and the edge [21,22] which is described by the Poisson model. Note also that a structure can have more than one edge component. The main aim for the design of the test structure is to reflect all critical structure issues which occur in the products. In other words the structure should be product relevant. For example, in a digital circuit with millions of MOS transistors an array of parallel connected MOS transistors is the ideal test structure assessing edge and area components of the gate dielectric simultaneously. The layout of a large structure in the scribe line has the disadvantage of generating a long structure which easily measures a few mm. For this reason the design and layout must be optimized to minimize the series resistance of the interconnects to the stress terminals [23]. Due to the limited space in the scribe line a maximum practical area of a fWLR test structure is approximately 10^{-3} cm^2. For thin gate oxide layers (approx. 3 nm and below) the maximum usable oxide test area is additionally limited from direct tunneling currents with respect to the max. current available from the test equipment. Different dielectrics are integrated in one integrated circuit and must be monitored, such as: MOS gate oxide or new high-k dielectrics, metal–insulator–metal (MIM) capacitors [24,25], polysilicon-oxide-polysilicon capacitors [26], intermetal dielectric (IMD) [27], stacked dielectrics [4], dielectrics for non-volatile memory cells [28] and tunnel oxides.

The dielectric reliability during fWLR monitoring on productive wafers is usually either measured with a ramped voltage stress (RVS) or with an exponentially ramped current stress (ERCS) [29,30] because of the short measurement times in the range of approximately 5–15 s. Both fast reliability stress methods can be adapted to all different dielectric materials, dielectric thicknesses, test structure sizes and types. They are documented in a JEDEC-standard [31] and can detect the catastrophic hard breakdown. In a voltage ramp a large current increase is expected in case of hard breakdown while during a current ramp a voltage drop is expected when the electrodes are shorted by a breakdown event. In case of the detection of soft breakdowns for thin gate oxide layers the breakdown control mechanism of the ramped stress must be modified which is described in a later subsection.

The most difficult part of a ramped oxide reliability test is the reliable automated breakdown detection. A current ramp offers the big advantage that a voltage drop is easy to detect compared to a current increase for a voltage ramp, especially in case of large direct tunneling currents. For a voltage stress usually a current increase (larger than 10 times [41]) indicates the hard breakdown event. This circumstance limits a voltage stress to smaller allowed tunnel currents than a corresponding current stress. Therefore, the voltage ramp restricts the testable oxide area for thin oxides. With the current ramp of [32] a 2.2 nm gate oxide area of 10^{-4} cm^2 can be reliably monitored without any problems using a SMU maximum current of 0.1 A.

A JEDEC conform ERCS had been first introduced to fWLR by Kamolz [33]. This stress measurement had the disadvantage of a very limited resolution of breakdown events in the low voltage regime. A newer more sophisticated method includes a short voltage ramp before the ERCS [32]. The stress sequence is schematically illustrated in Fig. 3. It consists of an integrity test at a low voltage (e.g. operating voltage), an initial voltage ramp and the exponential current ramp followed by two breakdown verification measurements. At each step of the current ramp approximately 10 voltage readings are taken. These readings are then used for a noise calculation in order to detect soft breakdown. In any case the stress is terminated after hard breakdown or reaching a maximum current density or maximum electric field. The measurement time for the detection of a hard breakdown of a gate oxide is less than 5 s.

4.1. Thin gate oxide breakdown and soft breakdown detection

Two different methods can be applied for the detection of thin gate oxide breakdown events:

- introduction of low bias integrity steps into the ramp,
- the assessment of the noise of the recorded measurement points.

First the RVS will be discussed: for the monitoring of breakdown events for thin gate oxides a JEDEC conform linear RVS is reported which additionally consists of low bias steps (at typically 0.1 MV/cm) as it is schematically illustrated in Fig. 4 [34,35]. This pulsed ramp had been first proposed by Hallberg [36] and Heimann [37]. In case of using these low voltage steps a breakdown event can be detected due to a large current increase at the low voltage level as it can be seen in Fig. 5 by the sudden jump in current while I_{use} is generally at 10^{-9} A [38]. The stress sequence of Fig. 4 has much longer stress times than a simple RVS due to the additional measurements. Also such a low voltage step is technically only feasible for a RVS but not for an ERCS.

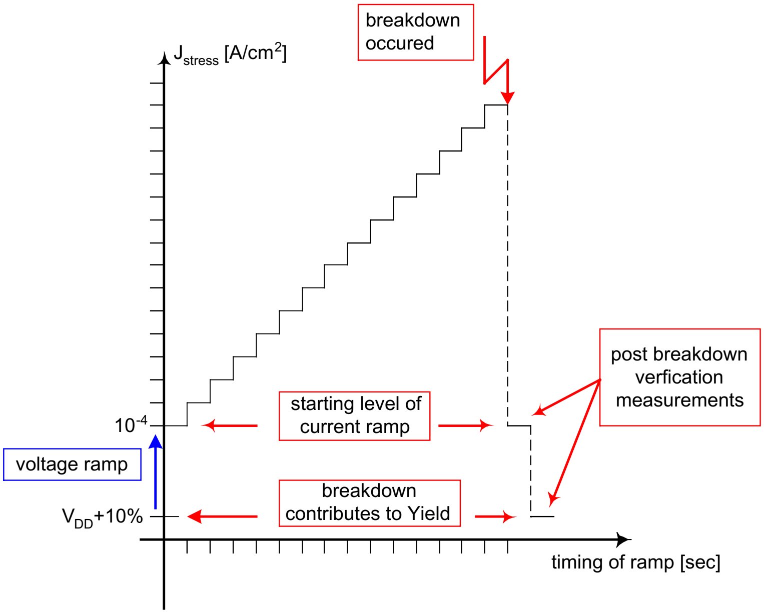
J_stress [A/cm²]
breakdown occured
post breakdown verfication measurements
starting level of current ramp
10⁻⁴
voltage ramp
V_DD+10%
breakdown contributes to Yield
timing of ramp [sec]

Fig. 3. Schematic illustration of an ERCS with a preceding voltage ramp and integrity and verification measurements [32].

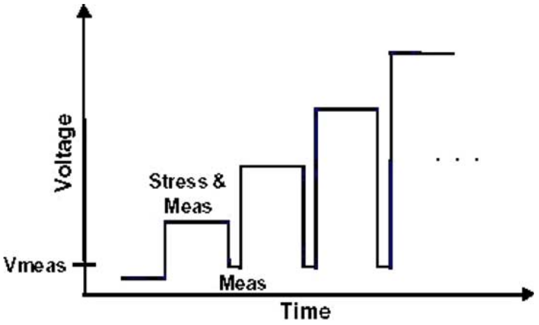

Fig. 4. Schematic illustration of a RVS with low voltage steps for the integrity measurements of the gate oxide leakage current.

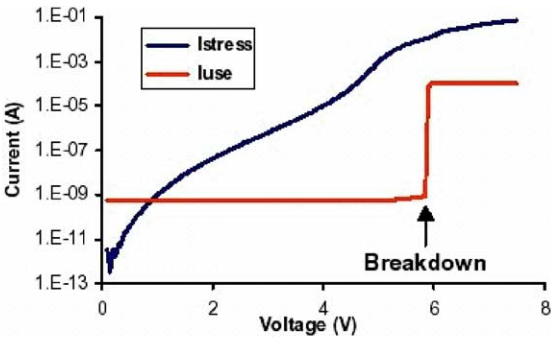

Fig. 5. Recorded data from RVS with low voltage integrity steps where the usual gate oxide leakage is at 10^{-9} A and increases suddenly orders of magnitudes in case of a breakdown [38].

Therefore, this method for soft breakdown detection seems less suitable for a fWLR monitoring when very short stress times are desired.

The detection of soft breakdown from noise increase is widely used and also manifested in a JEDEC standard [39–41]. This method is preferred for the fWLR monitoring stress sequence since it does not include additional time consuming measurements. It can be part of an ERCS and will be discussed in the following paragraphs. The fast noise detection method of reference [40] which requires five consecutive voltage readings has been slightly adjusted to get more reliable results. The noise is no longer calculated from the absolute value of the voltage readings but from the difference (slope) between two consecutive voltage readings. This offers the advantage of not being sensitive to any trapping effects (continuous voltage changes—increase or decrease) during a stress step of the current ramp. However, using the method calculating the noise from [40] the absolute value would indicate falsely an increased noise level during charge trapping. The noise of the slope method is calculated using Eq. (1) and requires six voltage readings at the same step of the ramp as input (V_i)

$$\mathrm{Noise}_{\mathrm{slope}} = \mathrm{ABS}\left[\frac{\sum_{i=1}^{5}(V_{i-1}-V_i)^2}{5} - \left(\frac{\sum_{i=1}^{5}(V_{i-1}-V_i)}{5}\right)^2\right] \tag{1}$$

In the equation V_i are the voltage readings per stress step of the current ramp. Please, note that the units of the noise in Eq. (1) are V^2.

In Fig. 6 the measurement raw data of the current ramp of [32] are displayed. The triangles forming an almost straight line represent the exponentially ramped current up to 3 mA and are plotted versus the number of recorded measurement points. The *x*-symbols are the corresponding measured voltages where it can be clearly observed at about measurement point 490 that the voltage decreases the first time. This event is a soft breakdown which does not trigger the standard hard breakdown criterion of a 5–10% voltage drop in an ERCS [30]. At around measurement point 590 a hard breakdown is recorded associated by a sudden drop of the voltage of nearly 1 V. The dashed noisy line in Fig. 6 are the resulting calculated voltage noise values using (Eq. (1)). A noise criterion for the detection of a soft breakdown is set to 10^{-4} indicated by the horizontal dashed line. In case of the first voltage decrease the calculated noise exceeds the criterion of 10^{-4} and clearly indicates the first breakdown event.

In Fig. 7 several Weibull distributions of charge to breakdown (Q_{bd}) from ERCS are presented emphasizing the need of soft breakdown detection and plotting the first breakdown event during an ERCS. The open triangles represent only automatically detected hard breakdowns with an underlying breakdown criterion of a 5% voltage drop during the ERCS. Also a sample is included which nearly had reached the Q_{bd} compliance of 50 C/cm^2. The open circles show the Q_{bd} of samples at automatic soft breakdown detection. Note, that in this case nearly all transistor arrays experienced a soft breakdown event. In order to plot the first breakdown event the minimum Q_{bd} of both is used and plotted as solid line which coincides with the soft breakdown distribution. It is clear from Fig. 7 that the Q_{bd} values of a hard breakdown event are misleading and overestimate the reliability of the gate oxide.

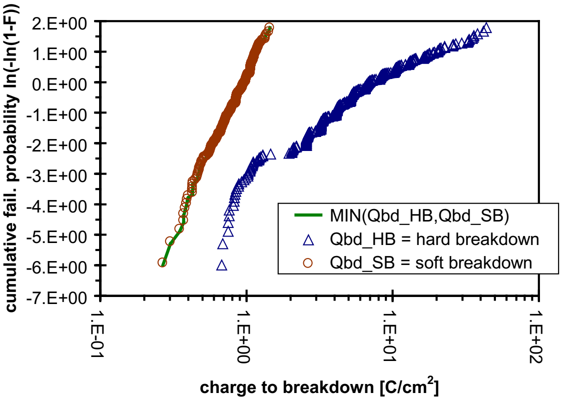

Fig. 7. Charge to breakdown (Q_{bd}) of hard and soft breakdowns measured during an ERCS displayed in a Weibull plot.

Fig. 8 presents Weibull distributions of voltage to breakdown (V_{bd}) of the same gate oxide as in Fig. 7. The open triangles represent automatically detected hard breakdowns while the open circles show the V_{bd} values of the soft breakdown. Also the first breakdown event is displayed which could either be the soft or hard breakdown event for the final distribution. If a soft breakdown occurs then the V_{bd} of the soft breakdown is plotted and in the other case the V_{bd} of the hard breakdown is used. Fig 8 indicates that the V_{bd} values of a hard breakdown event underestimate the reliability of a gate oxide because of breakdowns below the soft breakdown voltage and also due to the wide bimodal distribution which is not anticipated. A detailed description of the reason for the bimodal distribution is

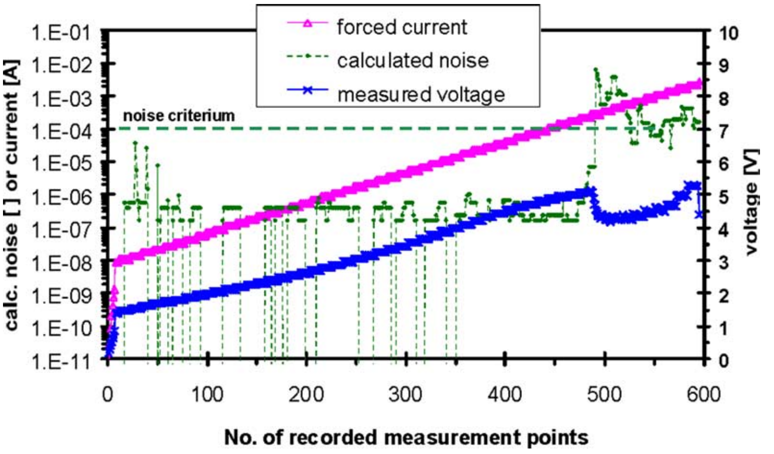

Fig. 6. Voltage readings, injected currents, and calculated noise versus the no. of recorded measurement points for a 2.2 nm PMOS oxide.

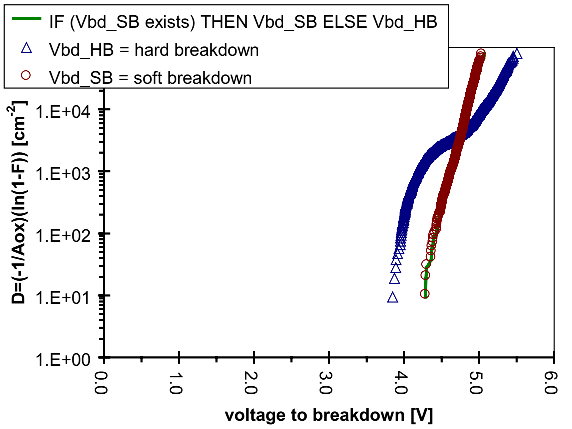

Fig. 8. Voltage to breakdown (V_{bd}) of hard and soft breakdowns recorded in an ERCS plotted in a Weibull diagram using a defect density scale on the y-axis.

given in [32]. Please note that in Fig. 8 the V_{bd} values are tightly distributed and show only intrinsic characteristics which can be easily fitted by a straight line in the Weibull diagram [15,29].

Additionally the defect density is introduced in Fig. 8 on the y-axis. The defect density, D, is calculated from Eq. (2), where A_{ox} is the oxide area of the test structure and $F = i/N$ the cumulative failed portion of the sample [29]. The defect density is independent of the area and is a measure of the dielectric quality

$$D = \left(\frac{-1}{A_{ox}}\right) \times (\ln(1-F)). \tag{2}$$

4.2. Dielectric data assessment

Two essential parameters are evaluated from a Weibull plot of a dielectric stress measurement and then monitored over time and many lots in a control card

1. a point representing the intrinsic reliability at 63.2% (intersection of distribution with dotted line in Fig. 9), optionally other percentages can also be extracted such as the 50% value since the intrinsic branch is tightly distributed;
2. a point which characterizes the extrinsic portion of the Weibull distribution (intersection of distribution with dashed line in Fig. 9).

Different methods are reported to extract a characteristic point from the extrinsic part of the distribution. A representative value can be determined from the separation point between the intrinsic and extrinsic branch. The separation point is located at the intersection of the distribution with the dashed horizontal line in Fig. 9. It is basically the onset of extrinsic behavior. In order to be able to compare the extracted values from various Weibull distributions it is necessary to plot the defect density on the y-axis as it has been shown in Fig. 8. The defect density is a unique parameter which is independent of the stress bias level, the dielectric area, and the type of reliability stress [29,42]. This defect density can also be compared directly with the targeted defect density for the integrated circuit. It has to be kept in mind that a targeted IC defect density of gate oxide is usually below 5 cm^{-2}. Assuming a realistic maximum dielectric test structure area of 10^{-4} cm^2 in the scribe line, 2000 transistor arrays would be required to monitor the defect density of 5 cm^{-2} with the point at the lowest

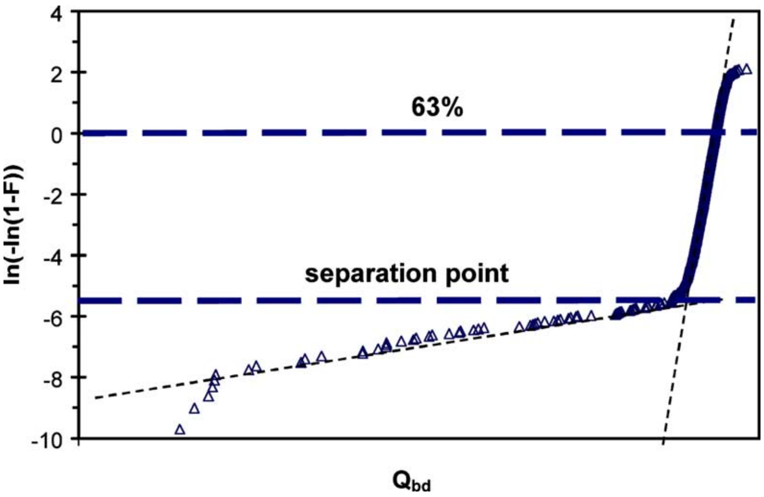

Fig. 9. Weibull distribution of 16,000 stressed transistor arrays of more than 60 lots with distinct extrinsic and intrinsic branches. Please note that the lowest defect related Q_{bd} deviate from the line due to the limited measurement resolution.

cumulative failure. Still this result would be not statistically relevant. It can be concluded that a much larger number of test structures must be measured than that which is usually available on one lot. Of course the situation is worse for a single wafer if one wants to draw a conclusion for the extrinsic dielectric reliability with respect to the targeted defect density. This underlines the statement which had been made in the earlier Section 2 regarding Table 1 that neither a wafer scrapping nor a lot scrapping makes sense in case of defect density monitoring unless the dielectric reliability is a catastrophe. In order to get some statement with respect to the product target dielectric defect density data can be cumulated over many lots [43]. Subsequently, an average defect density can be reported in a control card which has a good resolution below the targeted defect density. The cumulated data of Fig. 9 can serve as an example for such an assessment strategy.

5. Plasma induced damage

For productive fWLR monitoring of plasma induced damage (PID) usually a MOS transistor is employed, which has a much bigger structure, a so-called antenna, attached to the gate electrode. In some cases also MOS capacitors with antenna structures are reported [44]. But it has been shown over the last years in the literature that a MOS transistor as a PID detection device for fWLR is beneficial, especially for a detailed characterization sequence [45,46]. Consequently in this section only fWLR monitoring for PID with MOS transistors is discussed.

Typical antennas are 500–5000 times larger than the active gate oxide area. The ratio between antenna area and gate oxide area is called antenna ratio (AR). The test modules in the scribe line are often restricted to a set of few antenna test structures due to the limited space because of the large antenna area required. The typical worst case antenna is structured with minimum finger width and minimum spacing to address high antenna area ratios as well as high perimeter and shading effects [47] (see Fig. 10). Antennas for process monitoring can consist of single polysilicon, aluminum or copper metal layers and/or arrays of vias and contacts. The AR is much larger than the maximum AR allowed in the product. This magnification of the charging damage guaranties a certain safety margin or an early warning before a product relevant PID is detected [48]. For a multilevel metallization scheme stacked antennas are introduced where antennas of different layers are connected to one MOS gate in order to save test structure space. A stacked antenna signals as well as a single antenna the PID and indicates the group of process steps (e.g. FEOL = front end of line: poly-Si, contact) which possibly introduce a reliability risk. But a more detailed investigation is required in order to find out which specific layer had been affected. It is recommended to address each process level, from poly to the last-but-one metal level.

A pitfall with the design of an antenna and the corresponding AR is that during processing (especially plasma etching or plasma deposition) the actual real

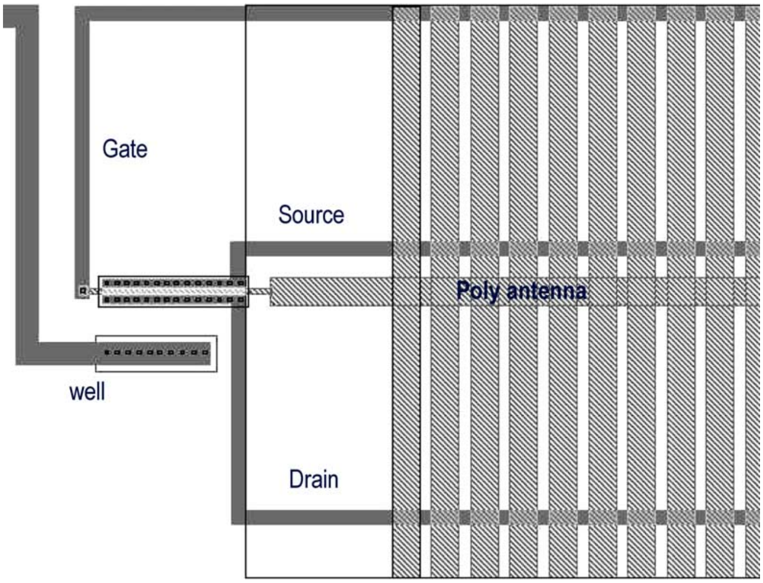

Fig. 10. Schematic of poly silicon antenna finger structure connected to the gate of a standard MOS transistor.

antenna can be much bigger than the designed antenna which is directly connected to the gate terminal. This aspect is described in detail in [49]. Another tricky point is that for avoiding the damage the maximum allowed AR is not always the correct measure with respect to the size of the active gate oxide area. When the active gate oxide area is large then the AR which could introduce PID becomes smaller than for a MOS transistor with minimum feature size [50]. In other words, AR is gate oxide area dependent. E.g. for large design conform transistor arrays with AR in the range of 2–5 a plasma damage can be observed.

Protective diodes or other schemes [49] should be used to avoid unwanted charging effects from bond-pads or connection lines especially on the reference device (MOSFET without designed antenna). For the MOS devices attached to the antennas only one single oxide thickness is used and an oxide area of a few μm^2 is preferred, to avoid any influence of gate oxide extrinsic defects and to minimize the required antenna area. Usually a medium thickness in the range of 4–5 nm is most susceptible to charging [51]. For monitoring purpose nMOS and/or pMOS transistors have been used, with the pMOS generally being more sensitive.

From this initial discussion it is clear that the main benefit of a PID monitor included in fWLR is that the process step(s) can be identified which cause the damage and consequently the processing tools can be adjusted. Due to a generally missing quantification of the PID it cannot be said exactly what the real damage will be on the product. But when gate oxide integrity structures and hot carrier MOS transistors are not diode protected then the influence of PID impact can be estimated from the performance of those reliability tests. Therefore, a downgrading or scrapping of wafers or lots solely based on the results of a PID monitor is not recommended.

5.1. Measurement sequence

Different types of stress measurements and types of parameters can be used to characterize the PID. In the literature three main characterization methodologies are reported:

1. Hot carrier stress test and the resulting transistor parameter drift [52].
2. Gate oxide reliability stress and the change in time to breakdown or breakdown strength [44,53].
3. Diagnostic gate bias stress and the corresponding transistor parameter drift [46,54].

A very short hot-carrier channel injection stress for fWLR is not as sensitive as a short gate bias Fowler–Nordheim stress to reveal PID latent damage because of the differences in stressed MOS transistor area. The applied fWLR stress should be very fast since being part of an in-line monitoring procedure. Therefore, any constant gate oxide stress (such as TDDB) measuring the time to breakdown is also not suitable. It has proven useful to perform a gate bias constant current stress, subsequently called diagnostic stress, and record the transistor parameter drifts.

Typically, as transistor parameters the gate oxide leakage I_G, the threshold voltage V_t, the transistor transconductance g_m, and/or the saturation current I_{ds} are measured. The most suitable transistor parameter for PID detection depends on the gate oxide thickness. As mentioned above the highest sensitivity of V_t to PID exists for 4–5 nm gate oxide thicknesses. Above those thicknesses V_t is also a suitable parameter for the PID characterization. However, below 4 nm the sensitivity of V_t on PID becomes low and instead the gate leakage current is more suitable for the characterization [55].

Deviations of the parameters measured on antenna devices with respect to those of the MOS reference device indicate the presence of PID. But performing the measurements on as-processed test structures may not reveal PID. Therefore, a gate oxide diagnostic stress is carried out on the test structures, in order to reveal the plasma induced damage that can be passivated during high-temperature process steps [56]. Fig. 11 shows as an example the cumulative distribution plots of the threshold voltages measured before and after the application of a constant-current stress (CCS) to antenna nMOS transistors. It can be clearly seen that only after the application of the CCS the latent PID can be observed. Fig. 11 also points out that the choice of the diagnostic stress is essential for the PID detection. After 50 mC/cm^2 all antenna devices and the reference device show extensive drift of V_t to negative values corresponding to a positive charge build up, induced by the CCS. After additional injection of 450 mC/cm^2 electron trapping is dominating and the M2 antenna of one of the two wafers clearly shows PID whereas no difference between reference and antennas can be seen on the other wafer. The correct stress conditions depend mainly on the oxide thickness, the gate oxide area, the PID mechanism and the type of dielectric [57].

The following measurement sequence is recommended to perform a correct in-line PID monitoring [54] in absence of a JEDEC standard for this reliability aspect:

1. Measurement of the gate oxide leakage current and the transistor parameters, V_t and, possibly, G_m and I_{ds}, in the antenna devices and in parallel also for the reference device. Depending on the process, some parameters could not be sensitive to PID [52] and in this case the characterization sequence can be shortened or modified accordingly.

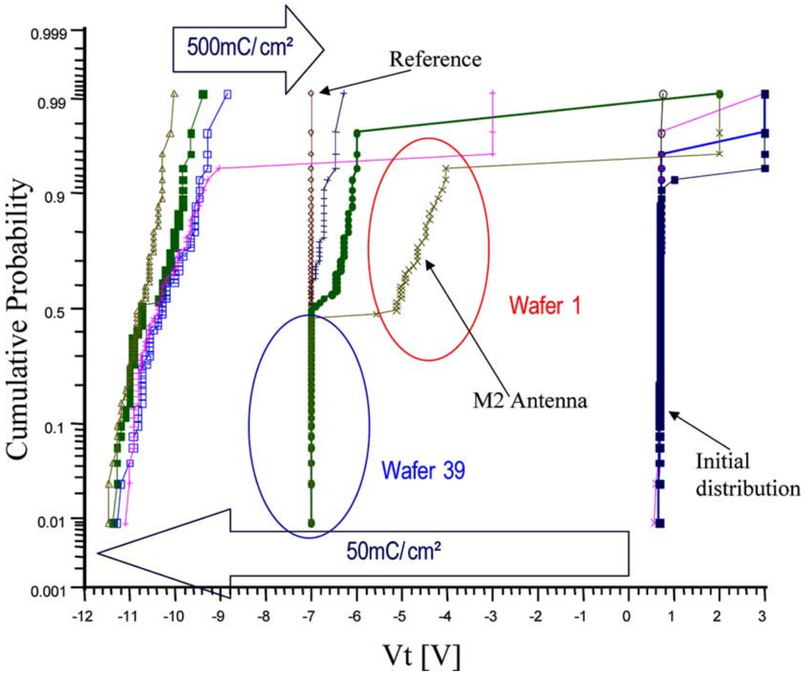

Fig. 11. Cumulative frequency plot; V_t distributions after injection of 0.05 and 0.5 C/cm^2 on 55 nm nMOS oxide of a reference transistor and M1, M2 and poly antennas.

2. Diagnostic stress in order to reveal plasma induced latent damage and to increase the probability of the characterization measurements to identify PID.
3. Repetition of the characterization measurements of step 1.

The first parameter measurement is the gate oxide leakage current. The correct choice of the characterization voltage is essential to identify PID. The characterization voltage should be chosen high enough. Fig. 12 presents the *I–V* characteristics of a 5.5 nm pMOS device with an antenna. By measuring the gate oxide current at the transistor supply voltage (2.5 V in the case reported in Fig. 12) the soft-breakdown (SB) would not be detected since the current is in the noise region and the resolution of an automatic in-line measurement system usually does not exceeds 1 to 10 pA. When the gate current is measured at a higher, properly tuned voltage (e.g. 4 V), the SBD events can be caught reliably. In conclusions it can be said that the gate leakage should be measured at an elevated voltage above the operating voltage.

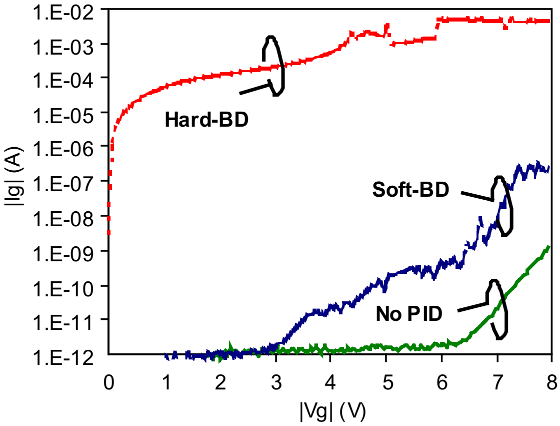

Fig. 12. Gate oxide *I–V* characteristics of scribe line antenna pMOS transistors with 5.5 nm gate oxide: to the right a device not affected by PID, a device showing soft-breakdown due to PID and the *I–V* curve (in the top of the diagram) of a device after the diagnostic stress had been carried out and the gate oxide had suffered hard-breakdown.

5.2. Analysis of the PID parameters

An in-line PID monitoring accumulates a huge quantity of recorded data which must be characterized, in order to draw correct conclusions about possible PID issues in the manufacturing process. One set of parameters is measured for each MOS transistor with different levels of antennas. In general, an automated procedure is necessary to organize the PID database and extract a manageable information. A methodology is described to extract two main PID parameters which are put in control/trend cards.

The assessment of the PID raw data must follow a fixed approach corresponding to the measurement sequence described above. It is essential that the assessment is carried out for the reference as well as for the antenna device. Four steps are necessary:

1. Characterization of gate oxide leakage.
2. Assessment of threshold voltage.
3. Identification of devices broken down during diagnostic stress.
4. Determination of drifts of transistor parameters before and after diagnostic stress and between antenna and reference device.

For steps 1 and 2 it is necessary to exclude outliers from the further analysis. For step 3 broken down devices must be excluded from further analysis in order to avoid misleading results [54]. An example is given in Fig. 13 where antenna structures before the diagnostic stress reveal significant deviations of V_t in comparison to the reference device (step 2 of the analysis approach). These structures must be excluded from further measurements and analysis.

As it has been demonstrated in [54] two parameters are very helpful in analyzing the drift data: The *Shift Mean* and the *sigma ratio*. The normalized averaged absolute shift of V_t for a wafer is called Shift Mean. It is the first important criterion for the PID assessment and is expressed in Eq. (3) where n is the total number of samples

$$\delta = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\Delta V_{t_i}}{V_{t_i}\_pre}\right|. \tag{3}$$

This Shift Mean parameter is calculated for antenna and reference devices and can be used for a comparison of PID even between different processes and oxide thicknesses. Additional PID shift indicated by significantly higher δ of the antenna devices compared to that of the reference devices is a measure for higher PID (or trap density [58]).

The broadening of the V_t parameter distributions caused by the revealing stress independent of the absolute V_t level the following can be expressed through parameter: the ratio between the standard deviation of the V_t distribution after stress and that before stress. It is a measure of the charging damage due to inhomogeneous processes and is described in Eq. (4) also called "sigma ratio" ρ:

$$\rho = \frac{\sigma(V_{t_i}\_post)}{\sigma(V_{t_i}\_pre)} \tag{4}$$

The sigma ratio is calculated for both the antenna device and the reference transistor and can be compared for different processes [59].

6. Interconnect fWLR tests

The high integration density of today's CMOS chips involves millions of wiring elements as metal lines, vias between metal levels and contacts to poly-silicon layers such as gate electrodes or to diffusions/implants (source/drain of devices). Hence these interconnects represent a reliability risk that needs careful assessment. Due to reduced cost, preparation and stress duration, which in

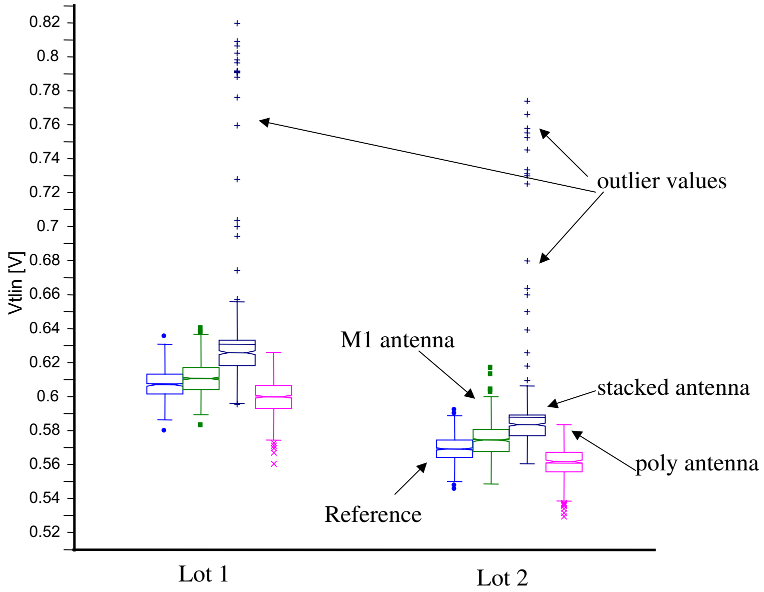

Fig. 13. Comparison of V_t distributions for two lots between antenna and reference transistors before diagnostic gate bias stress demonstrating initial high-flyer values.

turn allows higher sample size, a fast wafer level reliability test for electromigration (EM) has been proposed almost two decades ago with the introduction of the standard wafer level electromigration accelerated test (SWEAT) [60]. Besides the SWEAT the iso-thermal (constant temperature) [61,62] and the iso-current (constant current) reliability stress tests are also widely used for electromigration investigations. In many studies the different fWLR EM tests have been compared [63,64] extrapolated to lifetimes and correlated to package level stress [18,65–68] for aluminum and copper interconnects. The main focus of this section will be on the SWEAT which had been improved recently [69] and which is documented in an updated JEDEC standard [71]. The section concludes with a brief summary of contact and via fWLR tests.

6.1. The electromigration tests

The fundamental model behind the electromigration testing is Black's equation [17]

$$t_{\mathrm{tf}} = A \times J^{-n} \times \exp(E_{\mathrm{a}}/kT), \qquad (5)$$

where t_{tf} is the time to failure, A and n are material or structure related constants, J is the current density, E_{a} is the activation energy, k is Boltzmann's constant and T is the absolute temperature. This equation allows several ways to conduct the EM stress: the classical way is performed at constant temperature and current, for which the test structures are packaged and put in an oven at constant temperature, while a rather moderate current density is applied that does not cause significant additional heating. The test times are relatively long and cost as well as preparation time are considerable.

Shorter times are achieved by increasing the current density, which causes increasing temperatures in the metal lines due to Joule heating. This self-heating is exploited for fWLR and lead to the simple iso-current stress, which applies a high constant current. The Joule Heating causes the temperature to rise to an initial plateau, but especially towards the end of the stress when the metal line resistance increases due to the degradation caused by the stress the temperature rises further, i.e. the stress increases at the end of the stress. At the high current densities necessary for fWLR often a broadening of the distributions is observed if constant current stress is used, which is a result of locally varying Joule heating contributions. Although this test is straight forward and relatively easy to implement it requires exact knowledge of the temperature evolution to correct the distorted distribution.

An alternative is the iso-thermal stress, where the temperature is kept constant by means of controlling the power. The current is ramped up to the predetermined self-heated acceleration and as the resistance increases due to degradation the current is reduced, i.e. the stress decreases towards the end of the stress sequence. This requires a control loop that measures the temperature and adjusts the current. The implementation is at least tricky and not as straight forward as the constant current stress. Also reducing the stress over time increases the overall test time.

Therefore another idea is to leave the stress constant over time by adjusting the current based on measuring the voltage and calculating the temperature and the resistance simultaneously. This way the current and temperature change both but moderately during the stress while the stress acceleration remains constant. The acceleration is given by the right hand side of Black's equation excluding the constant A. This also requires a sophisticated control loop, which is not trivial to implement as the necessary revision of the JEDEC standard P119 has demonstrated [69]. The new SWEAT method that correlates well with package level stress results allows predictive fWLR monitoring for aluminum [65,70] and copper lines [18] as well.

6.2. The SWEAT

In preparation for the SWEAT test which is described in detail in [69,71] the desired acceleration needs to be determined, i.e. the constant A needs to be determined for the structure as well as for the material to be stressed. The acceleration is the ratio t_{tf}/A of Eq. (5), i.e. the acceleration increases as t_{tf} decreases. This allows for a wide range of stress conditions just by selecting t_{tf}. As soon as t_{tf} is set, the current density for an appropriate temperature can be searched assuming the activation energy (E_{a}) is known. Black's equation can be rearranged to yield the temperature for a given current density. A too high temperature may activate degradation mechanisms or non-uniformities in the heat distribution that are not product relevant, hence not desired. Once the adequate temperature and current density combination is determined for the chosen t_{tf} the stress preparation is complete. At the beginning of the stress the current is ramped up carefully preventing any overshoots to the determined current density. During the first part of the ramp up the thermal resistance R_{th} of the test structure is determined. R_{th} relates the temperature to the dissipated power ($P = I \times V$). The thermal resistance is the slope of the temperature versus power curve, where the temperature is calculated using the temperature coefficient of resistance (TCR). This relation is needed to determine the temperature of the test structure in the control loop, for which R_{th} has the advantage to be insensitive to any electromigration-induced resistance increase. For the correct R_{th} to be determined it should be taken in the range between 30 and 120 °C only, i.e. before any electromigration affects the measurement [69,71]. For aluminum lines that are stressed up to 300 °C this works fine. However, closer analysis reveals that the TCR and R_{th} are not linear over the entire temperature range. While for aluminum the deviation seems to be tolerable for copper that is stressed at temperatures as high as 600 °C the deviation becomes significant. Therefore, a correction factor is introduced [73] and R_{th} needs to be determined at stress temperature. This nonlinear temperature dependence is also manifested in a JEDEC standard [74]. During a SWEAT Black's equation is calculated with the actual values for current density and temperature continuously. As soon as the result equals the preset t_{tf} the ramp reached its target value. Now the control loop is activated and adjusts the current density if the in situ calculated t_{tf} leaves the ±1%-band around the preset t_{tf}. For this purpose Black's equation is rewritten in a way that t_{tf} is a function of current density only [69].

Fig. 14 illustrates the controlling of t_{tf} by means of the current density where the calculated t_{tf} is plotted versus the actual stress time. Whenever the t_{tf} is outside the control band (60 ± 1 s) the current density is adjusted, which moves the new t_{tf} closer to the preset target value. Fig. 14 also demonstrates that the preset t_{tf} (60 s) is a control parameter as a measure for acceleration only, while the time to reach the target degradation of the resistance is independent of it and in this case much longer (162 s). The observed peaks are probably induced by progressing electromigration. Note that in this case the boundary was set to ±1 s instead of 1%. The current reduction was less than 4% and the temperature increases less than 2% over the entire stress time in this case. The time to reach a certain resistance degradation, e.g. 20%, is plotted for all samples in a log-normal plot with all data falling on a straight line. An example for a log-normal distribution from a SWEAT fWLR stress is given in Fig. 15.

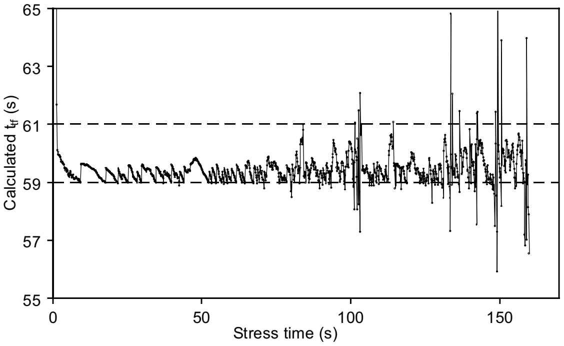

Fig. 14. The control parameter calculated t_{tf} as function of the actual stress time is well behaved even as electromigration progresses.

Using SWEAT it was successfully demonstrated that data from iso-current package level and SWEAT predict the same lifetime at operation conditions for aluminum [65] and copper [18] using Black's equation. In addition the failure mechanisms were analyzed by physical failure analysis and shown to be the same. Thus this offers the option to quantify the lifetime from fWLR SWEAT tests. There are some important pre-conditions: the constants in Black's equation need to be determined exactly and the temperature profile in the stressed line must be as uniform as possible. In addition the test structure needs to be designed for a test that uses Joule heating, i.e. it needs to consider the intended self-heating of the stressed line. In fact the limitations for SWEAT can be the test structure layout and too high temperatures.

For the test structure design it is necessary to be product relevant. Therefore, via terminated straight metal lines, as shown in Fig. 16 [18,72], are used to avoid

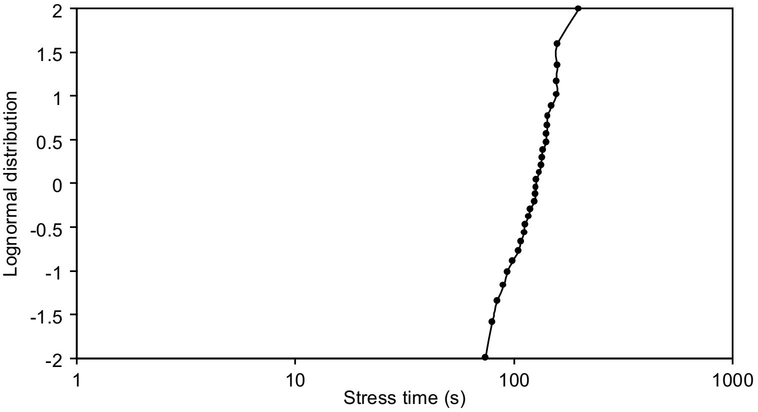

Fig. 15. Log-normal plot of t_{tf}, times to reach a certain degradation level. The data of the SWEAT is tightly distributed and can be approximated by a straight line.

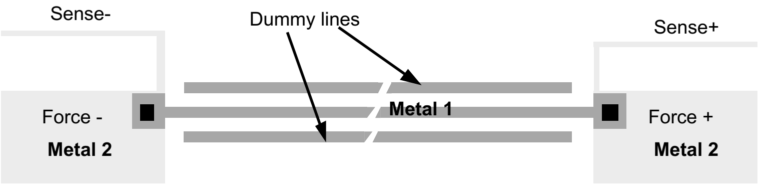

Fig. 16. Example for via-terminated M1 line as product relevant EM test structure. The electron flow from metal layer 2 to layer 1 can cause voiding near the via in the M1 line. [18,73].

the reservoir effect, e.g. from a connected pad. Via terminated lines have the advantage of shorter fail times than e.g. a NIST structure. Also the stressed metal lines do not include any line width variations (as a SWEAT structure [75]) to avoid heat sinks and to gain a very even temperature along the line. For electrical verification of the failure, its location and the detection of the possible failure mechanism additional sense lines are useful, e.g. sense lines above and below vias and other significant points. The temperature gradient between the stressed line and it's vicinity needs consideration for the temperature distribution. Also the temperature gradient between the stressed line and the supplying line as well as the sense lines needs to be taken into account for the layout. Another source for misleading results is overheating or failure of the supply lines, leading to shallower or even bimodal log-normal distributions. Dummy lines in parallel, as shown in Fig. 16, help to reduce temperature variations or can be used for cooling. Thermal simulations of the test structure at the intended stress conditions help checking whether the structure is feasible for the highly accelerated stress conditions of the self-heated EM test and support early detection of problems. An example is given in Fig. 17 where the supply line is too narrow and also heats up significantly. Thus the via is overheated and can cause an additional, undesired failure mode. The supply line can be broadened to avoid this effect. However, broadening it too much has the opposite effect (Fig. 18) and cools the line to be stressed, leading to too long times to failure due to an inhomogeneous temperature profile in the stressed line.

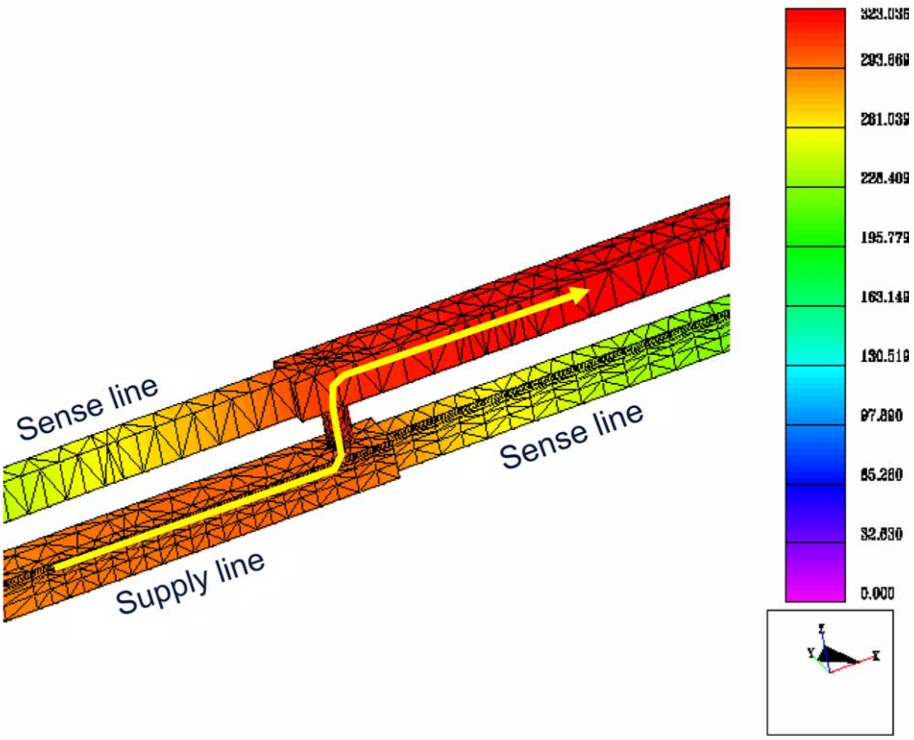

Fig. 17. FEM simulation for test structure design verification illustrating that the via will be overheated and the supply line heats up significantly, which is not intended. The arrow indicates the current direction.

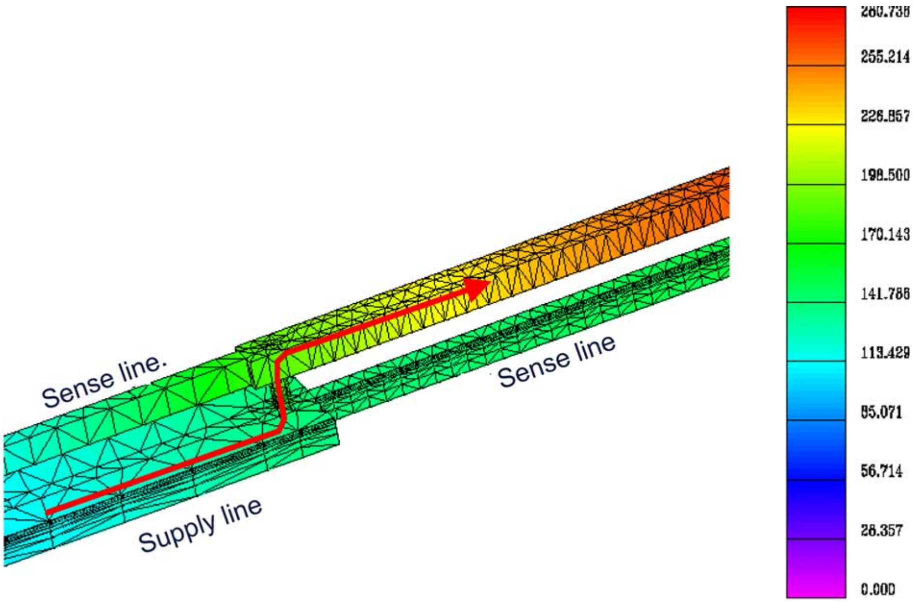

Fig. 18. FEM simulation for verification of test structure design. The supply line is too wide and acts as heat sink causing non-uniform heat distribution in the line that needs to be stressed. The arrow indicates the current direction.

SWEAT cannot be used for multiple stress lines connected in series or via chains because it is at least extremely difficult or not possible to determine the exact temperature in the stressed line. The iso-thermal stress using such structures has the same problem. Further challenges for newer technology generations are the low-K dielectrics in the metal stack that need careful evaluation whether the SWEAT is the suitable fWLR stress, because the thermal resistance is likely to change over the temperature range or may be in-stable at highly accelerated stress conditions.

For successful integration of a SWEAT in fWLR monitoring it is important to use careful screening. Thus the initial resistances of the stressed, supply and sense lines as well as the leakage currents to adjacent lines need to be recorded before and after stress. If the resulting parameters before stress are out of the expected specified resistance range the stress data are most likely not representative and should be taken out of the sample for a log-normal plot. The electrical detection and verification of the failure location was mentioned above already and is important to ensure that the intended stressed line did indeed fail. Regarding sampling the relatively long stress time (50–100 s) of a SWEAT allows a sample size, for example five chips per wafer. If SWEAT is performed on 5–10 wafers per lot a reasonable statistic is achieved for one lot supporting extraction of t_{50} and sigma for the control card. In case of bimodel distributions care must be taken that the t_{50} and the sigma is not determined from the entire distribution [76]. The earlier mode is the more important and can be separated from the longer t_{tf} values. Then t_{50} and sigma can be evaluated from the early mode only and reported in the control card. It might be necessary to increase the sampling when a bimodality is observed, especially when the earlier mode has a low probability to occur.

6.3. Contact and via fWLR tests

Another task of fWLR is to check the integrity of contacts and vias. Contact (or via) chains with a well adjusted amount of contacts (or vias) are measured. As depicted in Fig. 19, a contact chain for fWLR consists of contacts from first metal connected to n- or p-diffusions (or poly-Si). A test structure with two force and two sense lines ensures a precise resistance evaluation by means of four point measurements. Again accurate characterization of the test structure elements is essential. Initial resistance measurements of each connection

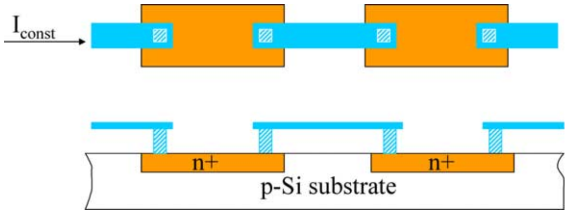

Fig. 19. Schematic top view and cross-section of a contact chain test structure. Contacts are positioned between short M1 lines and minimum n+ diffusion areas.

separately and if applicable reverse current measurements of involved pn-junctions allow detection of inconsistencies and data screening. After initial characterization at operation conditions a brief constant current stress is applied, which is typically less than ten seconds. Subsequently the characterization sequence as before stress is repeated and resistance drifts are determined. No commonly used model for lifetime prediction of the resistance drift is currently available in the literature. Therefore, the resistance after stress is typically plotted versus the resistance before stress for quick visual control as illustrated in Fig. 20. For data of the contact chain of Fig. 20 the constant current stress reveals two irregular not intended modes:

1. completely failed contacts with unacceptable high resistance (up to 5000 Ω) after stress,
2. contacts with initially high resistance which show improved resistance after stress.

The increased resistances of mode 1 indicate incompletely filled contact holes where the current is conducted along the barrier layer only. The reduction of the resistance of mode 2 is a result of breaking down insulating residuals in the contact hole. The advantage of the proposed diagnostic constant current stress is that appropriate corrective actions can be initiated according to the recorded results. For the control card reporting the resistance drift can be used. A ΔR_{50} and the sigma of the distribution are suitable for SPC control.

For via integrity similar tests and test structures are used as for contact integrity. The schematic cross-section is given as an example in Fig. 21, where the via connects two metal levels M1 and M2. Again the test structure should include connections for four point resistance measurements. In addition a poly-Si heater beneath the via chain allows local temperature acceleration. The stress sequence consists of:

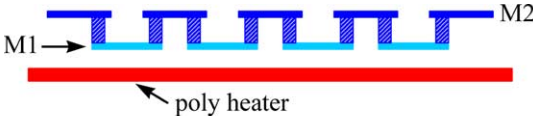

Fig. 21. Schematic cross-section of a self-heated via chain test structure. The poly-Si strip heats the structure above and accelerates the degradation.

1. the initial characterization of the test structure at room temperature,
2. the ramp up to temperature,
3. a brief constant current stress with a current low enough to not cause Joule heating,
4. cooling down phase,
5. the characterization after stress at room temperature.

For temperature control the power dissipated in the heater is kept constant during the stress. Simulations or an additional metal meander in one metal layer can be used to determine the temperature in the via chain.

7. Device reliability monitoring

Device reliability is another common reliability risk to be covered in integrated circuit processing and hence is an important part of every process qualification. Device parameter degradation occurs when the device is operated in a conducting (some voltage higher than the threshold voltage at the gate and the circuit operation specific voltage at the drain) [77–79] or a non-conducting

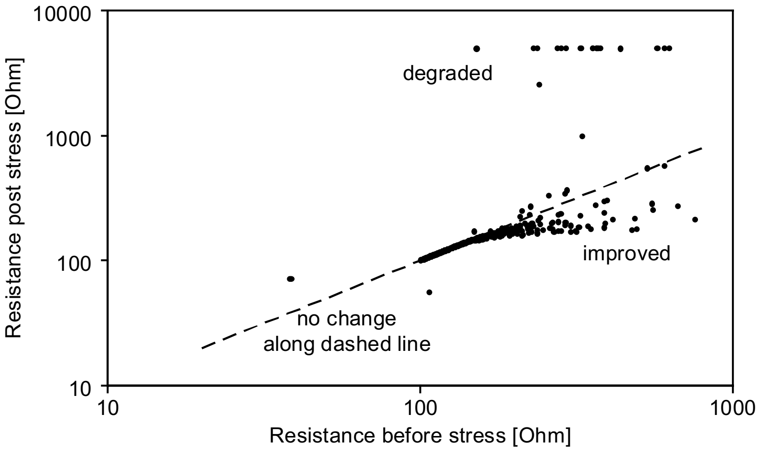

Fig. 20. Relation between structure resistance before and after constant current stress. The bad structures are due to open contacts in which the barrier layer conducts all current. The improved resistance of test structures reflects a breakdown of some residual insulating layer in the bottom of the contact hole.

mode (no gate bias, circuit operation specific voltage at the drain) [80]. Another condition is referred to as gate bias stress, for which the gate is biased, while well, source and drain are grounded. In some cases the temperature is also accelerated and it is called bias temperature stress (BTS) [81]. Depending on the device type, prevalent operation and the technology any of the three types could be the most critical condition. Therefore, during a qualification it is determined, which device at which condition is the most critical and must be monitored by means of fWLR monitoring.

Usually the two critical conditions are the conducting and non-conducting stress. The conducting hot carrier stress is performed at the gate and drain voltage combination that leads to the highest parameter degradation within the stress time [78,79]. For non-conducting device stress the drain voltage is chosen in a way that appreciable degradation occurs within reasonable time [80]. The relative parameter drift (in percent) is plotted as a function of time in a log–log plot with stress conditions as parameter. Increasing the stress conditions results in a similar degradation evolution over time [77,80], i.e. for different conditions the parameter drift as function of time yields parallel lines in a log–log plot as it is shown in Fig. 22. This is an indication that the degradation mechanism is consistent between stress conditions of process qualification and fWLR. By means of a model [77,80] the observed degradation is projected to operation conditions and the device reliability at the end of the specified product lifetime is estimated. Further details on device degradation can be found elsewhere [82,83]. Also other analysis methods [84] are reported than that in Fig. 22.

Particularly the conducting and the non-conducting stresses can be highly accelerated. Thus they are well suited for fWLR stresses and monitoring the technology device reliability beyond qualification. The degradation models of the process qualification can also be applied to the fWLR data and as a result lifetimes can be estimated from fWLR data. For both stresses the voltage applied to the drain is the main parameter accelerating the degradation. The limit for the acceleration is the so-called secondary avalanche breakdown when the drain voltage becomes too high. This point is characterized in the qualification and used later-on as limit for the fWLR conditions. Thus a short stress at highly accelerated stress conditions is the basis for fWLR monitoring. This stress takes approximately 10–20 s. The stress is performed on regular MOS transistor test structures. A typical fWLR stress and measurement sequence for an MOS transistor starts with the device characterization including gate and drain leakage and device parameter measurements (e.g. the threshold voltage). These parameters are used to determine a yield (integrity fails) before stress. If all parameters of the device under test are within the parameter specification limits the specified fWLR stress is applied followed by another characterization sequence of the device. The difference in the device parameters before and after stress represents the drift due to the stress. These drifts are plotted in the control card and compared to reference values from qualification hardware and other wafers from the processing time frame of the qualification. Typically the stress and measurement sequence is repeated on five chips per wafer. From each monitored lot 5–15 wafers are subjected to fWLR stresses.

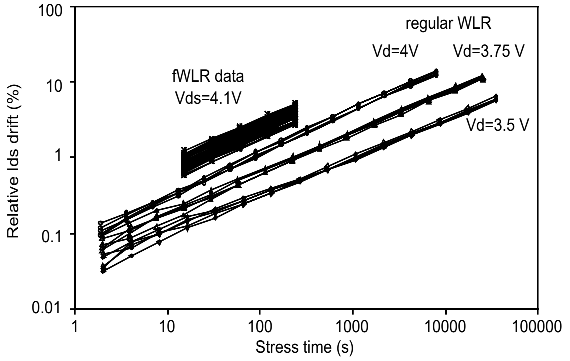

Fig. 22. Device I_{ds} drift as function of stress time for different stress voltages. Above the noise level the degradation curves from regular wafer level stress run parallel as the curves from fWLR at even higher acceleration. This indicates that degradation is driven by the same mechanism in all four cases.

Fig. 22 compares the degradation of a NFET under different conducting hot carrier stress conditions. ΔI_{ds} is plotted as a function of stress. Above the noise level ($\sim$0.2%) the curves run straight and parallel. At each drain voltage condition several devices were stressed to demonstrate the reproducibility. The highest stress conditions are used for fWLR. Several subsequent stresses were performed at different sites to demonstrate consistency with stresses at lower voltages. It is obvious that the slope is consistent with that at lower acceleration. The wider spread than for qualification stresses is the result of using a large wafer map and therefore showing the on-wafer variations.

For a robust process the highly accelerated but short stress may not be sufficient to reveal any degradation of the key parameters like the saturation drain current (I_{ds}) in forward or reverse direction. This is a positive result but to stay sensitive for a degradation the following measures can be applied:

- the stress duration can be extended,
- device parameter characterization before and after stress can be enhanced,
- fWLR can use additional transistor parameters that are more sensitive than the parameter of primary interest which correlate well with it.

One possible devices parameter which is more sensitive than e.g. I_{ds} is the analogue current I_A, which is measured at a gate voltage a few tenth of a volt above the threshold voltage. Using such indirect parameters helps to keep the stress time at a minimum while still detecting degradation above the noise level. Fig. 23 shows the drift of the analogue current as a function of the drift of the drain current in saturation. The triangle and the circle represent two different analogue currents. First of all the correlation between the drifts is excellent and supports the idea of the analogue current being the better choice for fWLR monitoring. As a second main conclusion from Fig. 23 it can be said that the drain current shifts merely three percent while the analogue current shows a 20% shift, which is much easier to detect and more accurate. Such an "indirect" parameter is especially beneficial for the signal to noise ratio, hence for the measurement accuracy.

The gate bias stress becomes more important at elevated temperatures. A well known example for a gate bias stress is the negative bias temperature stress (NBTS) on PFETs [85] that leads to serious instabilities also known as negative bias temperature instability (NBTI) [86]. The increased temperature necessary for a NBTS is a handicap for fWLR on a simple PMOS transistor, because the measurement sequence is usually carried out at a temperature slightly above room temperature. The required heating and cooling of the chuck for the complete stress and measurement sequence would take considerable time and in addition is not desired to be applied to the products on the same wafer. Therefore, structures that can be heated locally are used [87]. Such heater structures use poly silicon lines for resistive heating with metal lines on top to control the temperature. An example is described in another contribution of this issue [88], to which we refer for more details and results.

Locally heated structures are also well suited for mobile ion detection. In this case a transistor using a metal gate over thick isolation oxide or just a simple PMOS transistor combined with a poly-Si heater is useful. Again the temperature is controlled by resistance measurement of a metal line on top of the poly-Si heater. A typical stress sequence is shown in Fig. 24. Since the dependence of mobile ions on the electric field across the dielectric is weak, the temperature dominates the stress acceleration. Therefore, the stress is performed at temperatures in the range of 200–250 °C. A key element is to leave the bias on the gate until the poly heater has cooled down to room temperature. This allows freezing the ions at the location where they had moved during the stress. If this is not considered and the bias is taken off first the ions may redistribute and an erroneous or no V_t-shift is observed. Before and after the high temperature gate bias stress the transistor parameters are measured at room temperature. In Fig. 25 the V_t shifts of mobile ion stresses with different polarities are displayed. After

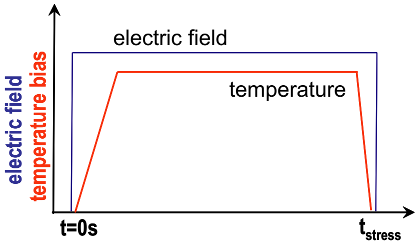

Fig. 24. Stress sequence for fWLR mobile ion testing using a poly-Si heater. Having the bias on the device under test all the time ensures that the ions are not redistributed.

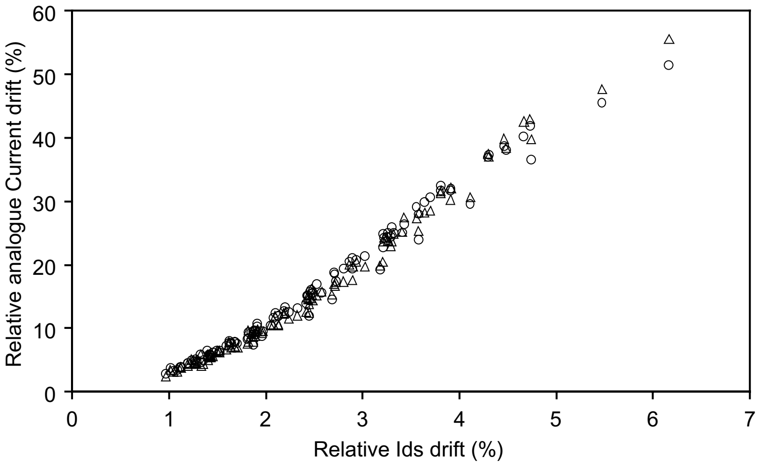

Fig. 23. Correlation between analogue current drift and drain current drift. The triangle and the circle represent two different analogue current; both correlate excellent with the drain current degradation.

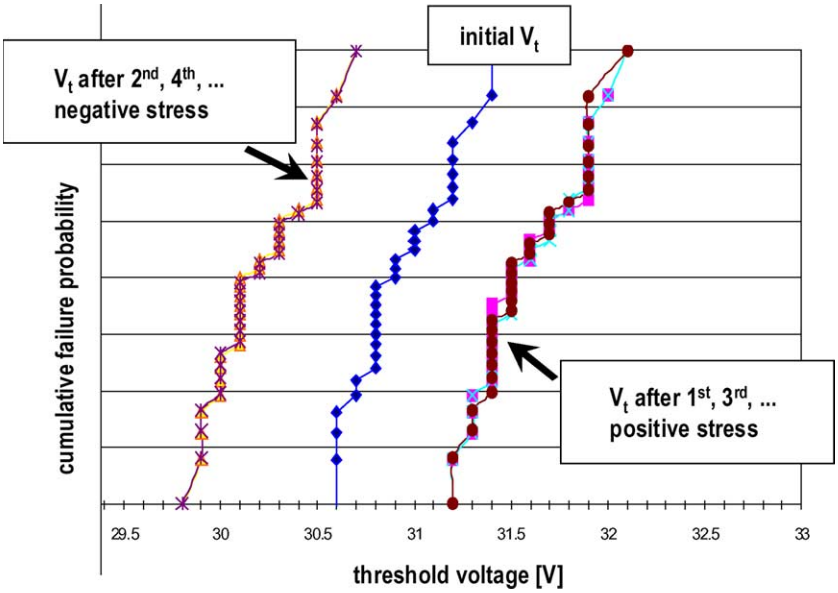

Fig. 25. V_t-distributions from fWLR mobile ion test. Depending on the gate bias the ions can be moved to the metal electrode or the silicon interface and back.

applying a positive voltage to the gate of a PFET the V_t shifts to higher threshold voltage, indicating that the ions are pushed down to the SiO_2–Si-interface. A second stress with negative gate bias moves the ions up to towards the metal electrode reducing the influence of the ions on the channel and hence the threshold voltage. The ions can be moved reproducibly and the original V_t distribution is restored by applying the temperature over a longer time without bias.

8. Conclusions

The use of fWLR monitoring on productive wafers is without doubt the essential tool to systematically demonstrate the current process reliability performance with respect to the targeted process reliability for integrated circuit fabrication after process qualification. Application of an SPC approach on fWLR data indicates if the process reliability is stable and offers the advantage to trigger corrective actions at an early state of IC production. For most of the reliability mechanisms highly accelerated tests have been reported and are commonly available for use in in-line measurements. fWLR monitoring has also shown its use for process development before process qualification. It can reduce the traditional measurement effort of long term or package level tests significantly. The benefit of fWLR monitoring is clearly recognized by the technical community and by now most of the semiconductor manufacturers have already implemented fWLR tools in their fabrication areas. The success stories reported in the literature indicate this trend towards fWLR.

In this introduction to fWLR monitoring it has been shown that only a careful and complete approach from the design and layout of the scribe line test structure, the set up of a sensitive stress and measurement sequence to the correct and exact analysis method will result in the maximum benefit and a correct interpretation. At the same time the sampling and frequency of fWLR monitoring is crucial to the detection of reliability violations. It became clear that a minimum sample size for reliability mechanisms is required and in some cases the testing of every single wafer with 1–3 samples is not sufficient to get a clear picture of the reliability performance. Surely not all angles and aspects of fWLR monitoring are discussed in this introduction but it highlights many pitfalls and areas of concern for the described test structures and reliability stresses. Also it has been indicated that in some fWLR areas more research must be done in order to optimize stresses or adapt existing stresses to new process options.

Acknowledgements

The authors thank the following colleagues for their contributions in the field of the development of fWLR monitoring: Alexander Ambatiello, Georg Antonin, Wilhelm Asam, Josef Fazekas, Jochen von Hagen, Werner Muth, Andreas Pietsch, and David Smeets. Additionally we acknowledge the detailed discussions and the critical review of the manuscript by Josef Fazekas, Werner Muth and Andreas Pietsch.

References

[1] Papp A, Bieringer F, Koch D, Kammer H, Pohle H, Schlemm A, et al. Use of test structures for a wafer level reliability monitoring. IEEE ICMTS 1996;9:267–71.

[2] Schafft H et al. Design of experiments and innovative measurement for BIR. In: International Wafer Level Reliability Workshop; 1991. p. 229–37.

[3] Schafft HA, Erhart DL, Gladden WK. Toward a building-in reliability approach. Microelectron Reliab 1997;37(1): 3–18.

[4] Madson G, Probst D, Rawlins L. Building reliability into an EPROM cell using in-line WLR monitors. IEEE IRW 1996:40–4.

[5] Crook DL. Evolution in VLSI reliability engineering. IEEE IRPS 1990:2–11.

[6] Turner T. Enhanced wafer level electromigration test structure for process reliability control. In: International Wafer Level Reliability Workshop; 1990. p. 115–24.

[7] Messick C. Implementation of wafer level quality program. In: International Wafer Level Reliability Workshop; 1990. p. 125–32.

[8] Garrad S. Production Implementation of a Practical WLR Program. IEEE IRW 1994:20–9.

[9] Yap KL, Yap HK, Tan YC, Lo KF, Karim MF, Manna I. Implementation of FWLR for process reliability monitoring. IEEE IRW 2001:94–6.

[10] Papp A, Bieringer F, Koch D, Kammer H, Kohlhase A, Lill A, et al. Implementation of a WLR-program into a production line. IEEE IRW 1995:49–54.

[11] Kanak Sarma, Mahmoud Bahrami, Kaizad Mistry. Wafer level reliability application to manufacturing of high performance microprocessor. IEEE IRW 1997:77–81.

[12] McPherson JW, Baglee DA. Acceleration factors for thin gate oxide stressing. IEEE IRPS 1985:1–5.

[13] Vollertsen RP. Thin dielectric reliability assessment for DRAM technology with deep trench storage node. Microelectron Reliab 2003;43:865–78.

[14] Suehle JS et al. Experimental investigation of the validity of TDDB voltage acceleration models. IEEE IRW 1993: 59–67.

[15] Hunter RW. The analysis of oxide reliability data. IEEE IRW 1998:114–34.

[16] JEDEC-standard, JESD 37, Standard for Lognormal Analysis of Uncensored Data and of Singly Right-Censored Data Utilizing the Persson and Rootzen Method; October 1992.

[17] Black JR. Mass transport of aluminum by momentum exchange with conducting electrons. In: Proceedings of the IEEE IRPS; 1967. p. 148–59.

[18] von Hagen J, Bauer R, Penka S, Pietsch A, Walter W, Zitzelsberger A. Extrapolation of highly accelerated electromigration tests on copper to operating conditions. IEEE IRW 2002:41–4.

[19] Snyder ES, et al. Self-stressing structures for wafer-level oxide breakdown to 200 MHz. IEEE IRW 1994:113–7.

[20] Dumin DJ et al. Test structures to investigate thin insulator dielectric wearout and breakdown. IEEE ICMTS 1991:61–7.

[21] Uchida H, Aikawa I, Hirashita N, Ajioka T. Enhanced degradation of oxide breakdown in the peripheral region by metallic contamination. IEEE IEDM 1990:405–8.

[22] Kerber M, Zeller C. Impact of oxide thinning at the LOCOS edge of MOS capacitors on constant current stress. In: Proceedings of the ESSDERC; 1989. p. 139–42.

[23] Pio F. Sheet resistance and layout effects in accelerated tests for dielectric reliability evaluation. Microelectron J 1996;27:675–85.

[24] Allers KH, Schwab R, Walter W, Schrenk M, Koerner H. Thermal and dielectric breakdown for Metal Insulator Metal Capacitors (MIMCAP) with tantalum pentoxide dielectric. IEEE IRW 2002:96–101.

[25] Martin A, von Hagen J, Fazekas J, Allers KH. Fast and reliable WLR monitoring methodology for assessing thick dielectric test structures integrated in the Kerf of product wafers. IEEE IRW 2002:83–7.

[26] Martin A et al. Correlation of lifetimes from CVS and RVS using the 1/e-model for thermally grown oxides on polysilicon. IEEE IRW 1994:106–12.

[27] Alers GB, Harm G, deFelipe TS. Wafer level testing of inter-line reliability in copper/low-k structures. IEEE IRW 2001:83–6.

[28] Bellafiore N et al. Thin oxide nitridation in N_2O by RTP for non-volatile memories. Microelectronics Journal 1993;24:453–58.

[29] Wolters DR et al. Dielectric breakdown in MOS devices, Part I: defect-related and intrinsic breakdown. Philips J Res 1985;40(3):115–36.

[30] Martin A et al. Dielectric reliability measurement methods: a review. Microelectron Reliab 1998;38(1):37–72.

[31] JEDEC Standard, JESD35-A, Procedure for Wafer-Level Testing of Thin Dielectrics; April 2001.

[32] Martin A et al. Ramped current stress for fast and reliable wafer level reliability monitoring of thin oxide reliability. Microelectron Reliab 2003;43(8):1215–20.

[33] Kamolz M. CSQ-Test: A special J-Ramp method approved for fast routine testing of thin dielectric films. In: WLR'91 (former IRW); 1991. p. 121–32.

[34] Martin A, et al. Correlation of lifetimes from CVS and RVS using the 1/e-model for thermally grown oxides on polysilicon. IEEE IRW 1994:106–12.

[35] Martin A, et al. A new oxide degradation mechanism for stresses in the Fowler–Nordheim tunneling regime. IEEE IRPS 1996:76.

[36] Hallberg Ö. NMOS voltage breakdown characteristics compared with accelerated life tests and field use data. IEEE IRPS 1981:28–33.

[37] Heimann AP. An operational definition for breakdown of thin thermal oxides of silicon. IEEE Trans Electron Dev 1983;30(10):1360–8.

[38] Snyder ES, et al. Detecting breakdowns in ultra-thin dielectrics using a voltage ramp. In: IEEE IRW'99; 1999. p. 118–23.

[39] Roussel P, Degraeve R, Van den Bosch G, Kaczer B, Groeseneken G. Accurate and robust noise-based trigger algorithm for soft breakdown detection in ultra thin oxides. In: Proceedings of the 39th IEEE International

Reliability Physics Symposium, IRPS 01, Orlando, FL; April/May 2001. p. 386–91.

[40] Alers GB, Weir BE, Frei MR, Monroe D. J-Ramp on sub-3 nm dielectrics: noise as a breakdown criterion. In: Proceedings of the 37th IEEE International Reliability Physics Symposium, IRPS 99, San Diego, CA; March 1999. pp. 410–3.

[41] JEDEC-standard, JESD92, Procedure for characterizing time dependent dielectric breakdown for ultra-thin gate dielectrics; August 2003.

[42] Martin A, et al. Method for the extraction of extrinsic data for oxide quality assessment. IEEE IRW 2000:187–8.

[43] Martin A, et al. The challenge to record correct fast WLR monitoring data from productive wafers and to set reasonable limits. In: Proceedings of IEEE IRPS; 2004:661–2.

[44] Eriguchi K et al. Quantitative evaluation of gate oxide damage during plasma processing using antenna structure capacitors. Jpn J Appl Phys 1994;33:83ff.

[45] Zhao J, et al. Investigation of charging damage induced V_t mismatch for submicron mixed-signal technology. IEEE IRPS 1996:33ff.

[46] Brozek T et al. Threshold voltage degradation in plasma damaged cmos transistors—role of electron and hole traps related to charging damage. Microelectron Reliab 1996; 36:1637ff.

[47] Hashimoto K. New phenomena of charge damage in plasma etching: heavy damage only through dense-line antenna. Jpn J Appl Phys 1993;32:6109–13.

[48] Cheung KP, et al. Plasma charging damage of ultra-thin oxides—the measurement dilemma. In: Proceedings of the 5th International Symposium on Plasma Process-Induced Damage (P2ID); 2000. p. 10–3.

[49] Paul Simon, et al. Antenna ratio definition for VSLI circuits. In: Proceedings of 4th International Symposium on Plasma Process-Induced Damage (P2ID); 1999. p. 16–20.

[50] Noguchi K, et al. A model for evaluating cumulative oxide damage from multiple plasma processes. IEEE IRPS 2000:364–9.

[51] Alavi M, et al. Effect of MOS device scaling on process induced gate charging. In: Proceedings of 2nd International Symposium on Plasma Process-Induced Damage (P2ID); 1997. p. 7–10.

[52] Hook T, et al. A comparison of hot electron and Fowler–Nordheim characterisation of charging events in a 0.5 μm CMOS technology. In: Proceedings of 1st International Symposium on Plasma Process-Induced Damage (P2ID); 1996. 164f.

[53] Hook T, et al. Detection of thin oxide (3.5 nm) dielectric degradation due to charging damage by rapid-ramp breakdown. In: Proceedings of the 38th Annual International Reliability Physics Symposium (IRPS); 2000. p. 377–88.

[54] Smeets D, Martin A, Fazekas J. WLR monitoring methodology for assessing charging damage on oxides thicker than 4 nm using antenna structures. IEEE IRW 2001:67–73.

[55] Cheung KP. Initial gate leakage in ultra thin SiO_2—the role of a brief stress. In: Proceedings of 8th International Symposium on Plasma Process-Induced Damage (P2ID); 2003. p. 122–5.

[56] Cheung KP. On the use of Fowler–Nordheim stress to reveal plasma-charging damage. In: Proceedings of 1st International Symposium on Plasma Process-Induced Damage (P2ID); 1996. p. 11–4.

[57] van den Bosch G, et al. Evaluation procedure for fast and realistic assessment of plasma charging damage in thin oxides. In: Proceedings of 7th International Symposium on Plasma Process-Induced Damage (P2ID); 2002. p. 37–40.

[58] Smeets D, et al. Quantifying charging damage gate oxides of antenna structures for WLR monitoring. Microelectron Reliab 2004;44:1245–50.

[59] Smeets D, et al. A general concept for monitoring plasma induced charging damage. In: Proceedings of 8th International Symposium on Plasma Process-Induced Damage (P2ID); 2003. p. 36–9.

[60] Root BJ, Turner T. Wafer level electromigration tests for production monitoring. IEEE IRPS 1985:100–7.

[61] Lee S-Y, Lai JB, Lee SC, Chu LH, Huang YS, Shiue RY, et al. Real case study for isothermal EM test as a process control methodology. In: Proceedings of the IEEE IRPS; 2001. p. 184–8.

[62] EIA/JEDEC-standard, EIA/JESD61, Isothermal Electromigration Test Procedure; April 1997.

[63] Menon SJ, von Hagen J, Head LM, Ellenwood CH, Schafft HA. Impact of test-structure design and test methods for electromigration testing. IEEE IRW Final Report; 1999, p. 46–53.

[64] Lee TC, Tibel D, Sullivan TD. Comparison of isothermal, constant current and SWEAT wafer level EM testing methods. In: Proceedings of the IEEE IRPS; 2001. p. 172–83.

[65] Zitzelsberger A, Pietsch A, von Hagen J. Electromigration testing on via line structures with a SWEAT method in comparison to standard package level tests. IEEE IRW Final Report; 2000. p. 57–60.

[66] Lepper M, Bauer R, Zitzelsberger AE. A correlation between highly accelerated wafer level & standard package level electromigration tests on deep sub-micron via-line structures. IEEE IRW Final Report; 2000. p. 70–73.

[67] Ryu C, Tsai T-L, Rogers A, Jesse C, Brozek T, Zarr D, et al. Experimental comparison of Wafer Level Reliability (WLR) and packaged electromigration tests. In: Proceedings of the IEEE IRPS; 2001. p. 189–93.

[68] Tibel D, Sullivan TD. Comparison of via/line package level vs. wafer level results. In: Proceedings of the IEEE IRPS; 2001. p. 194–9.

[69] v.Hagen J, Antonin G, Fazekas J, Head L, Schafft H. New SWEAT method for fast, accurate and stable electromigration testing on wafer level. IEEE IRW Final Report; 2000. p. 85–9.

[70] Yap KL, Lim BH, Yap HK, Tan YC, Lo KF. Real case study of SWEAT EM test on via/line structure as process reliability monitor methodology. IEEE IRW 2002:165–7.

[71] JEDEC-standard, JEP119-A, A Procedure for Performing SWEAT; August 2003.

[72] JEDEC-standard, JESD 87, Standard Test Structures for Reliability Assessment of AlCu Metallization with Barrier Materials; July 2001.

[73] v.Hagen J, Schafft H. Temperature determination methods on copper material for highly accelerated electromigration tests (e.g. SWEAT). IRW Final Report; 2002. p. 45–9.

[74] JEDEC-standard, JESD33-B, Standard method for measuring and using the temperature coefficient of resistance to determine the temperature of a metallization line; February 2004.

[75] Turner T. Enhanced wafer level electromigration test structure for process reliability control. WLR Final Report; 1990. p. 115–24.

[76] Fischer AH, et al. Experimental data and statistical models for bimodal EM failures. IEEE IRPS 2000:359–63.

[77] Takeda E, Suzuki N. An empirical model for device degradation due to hot-carrier injection. Electron Dev Lett 1983;4:111–3.

[78] JEDEC-standard, JESD 28-A, Procedure for measuring N-channel MOSFET hot-carrier-induced degradation under DC stress; December 2001.

[79] EIA/JEDEC-standard, EIA/JESD 60, A procedure for measuring P-channel MOSFET hot-carrier-induced degradation at maximum gate current under DC stress; April 1997.

[80] Muehlhoff A. An extrapolation model for lifetime prediction for off-state-degradation of MOS-FETs. Microelectron Reliab 2001;41:1289–93.

[81] Denais M, et al. Interface traps and oxide traps creation under NBTI and PBTI in advanced CMOS technology with a 2 nm gate-oxide. IEEE IRW 2003:1–6.

[82] Bravaix A. Hot-carrier degradation evolution in deep submicrometer CMOS technologies. IEEE IRW Final Report; 1999. p. 174–83.

[83] LaRosa G, Rauch S, Guarin F. New phenomena in device reliability physics of advanced CMOS submicron technologies. IEEE IRPS Tutorial; 2001.

[84] JEDEC-standard, JESD 28-1, N-Channel MOSFET hot carrier data analysis; September 2001.

[85] Schlünder C, Brederlow R, Wieczorek P, Dahl C, Holz J, Röhner M, et al. Trapping mechanisms in negative bias temperature stressed p-MOSFETs. Microelectron Reliab 1999;39:821–6.

[86] LaRosa G. NBTI challenges in PMOSFETs of advanced CMOS technologies. IEEE IRPS Tutorial; 2003.

[87] Muth W, Martin A, von Hagen J, Smeets D, Fazekas J. Polysilicon resistive heated scribe lane test structure for productive wafer level reliability monitoring of NBTI. IEEE ICMTS 2003:35–41.

[88] Muth W, Walter W. Bias temperature instability assessment of n- and p- channel MOS transistors using a polysilicon resistive heated scribe lane test structure. Microelectron Reliab 2004;44:1251–62.