A Novel In-Process Wafer-Level Screening Technique for CMOS Devices

I. Yoshii, K. Hama *, H. Hazama, H. Kamijo, and Y. Ozawa

Toshiba Corporation
and
* Toshiba Microelectronics Corporation

72, Horikawa-cho, Saiwai-ku, Kawasaki, 210 Japan
Phone: 44-549-3528 Fax: 44-549-3213

Abstract— We have developed a novel in-process wafer-level screening technique to eliminate infant mortality of CMOS device due to gate oxide defects. Using this technique, it is possible to stress all of gate oxides simultaneously at an arbitrary high voltage for both n-channel and p-channel transistors. This paer describes the detailes of the sceening method and its application to the standard 0.8μm CMOS logic technology. The result shows that eraly TDDB failures are significantly reduced by this technique.

Keywords— **CMOS, reliability, TDDB, defect, screening, burn-in, early failure.**

I. Introduction

One of the most important issues for CMOS devices is the reliability to the time-dependent dielectric breakdown (TDDB) caused by gate oxide defects. In order to produce MOS devices with no oxide defects, extensive work has been made; such as decrease in the defect density by fabrication process improvement, removal of the devices with oxide film defects by screening.

To reduce or eliminate oxide defects of MOS devices, some attempts to process technologies, such as use of high-quality wafer [1], improvement of oxidation process [3], and addition of nitrogen to gate oxide [2] have been proposed. However, there has been no production technique established which is able to remove the gate oxide defects thoroughly and assure TDDB reliability.

On the other hand, from the point of view of screening, the burn-in of the finished products is widely used. This method stresses the gate oxide through external terminals at the end of the line at a high temperature, so that it is unable to activate and stress every gate oxide of the device, especially for logic, within the limited time. Therefore, the burn-in hardly makes it possible to assure the gate oxide reliability completely. Moreover, since the transistor breakdown voltage, e.g., junction breakdown voltage and snapback voltage, is generally lower than or comparable with the gate oxide breakdown voltage, the voltage higher than the transistor breakdown voltage cannot be applied for the burn-ih stress. This sets a limit to increasing the stress voltage to increase acceleration for enhancing the effectiveness of burn-in. Furthermore, burn-in has an essential problem that the stress is applied even to non-defective gate oxides. This results in reducing the life time of surviving gate oxides since no damage recovery process is available after burn-in.

These disadvantages in burn-in is originated from the fact that the stress is applied after each transistor on a die is connected by metallization to form circuits. Therefore, if screening is performed prior to metal interconnection and the stress is applied only to the gate oxides with drain and source free from the stress voltage, the above-mentioned difficulties can be resolved; there is no restriction to the stress voltage due to the transistor breakdown voltage, and the damage to the gate oxide caused by stress can be recovered by high-temperature annealing before metal interconnection.

The idea of the screening during fabrication process has been discussed by King et al. [4] in 1994. Their discussion, however, is very preliminary and no detailed implementation of an in-process wafer-level screening technique to the actual CMOS process technology has been shown. In this paper, we present a novel in-process wafer-level screening technique which can be easily inserted to any CMOS process technologies, and demonstrate the effectiveness of this technique integrated into the 0.8 μm double-metal CMOS technology.

II. Basic ideas behind the in-process wafer-level screening

As shown in Figure 1, the oxide breakdown, which results in the reliability failure of devices, can be classified roughly into two modes (B and C modes). The C-mode breakdown corresponds to the intrinsic breakdown of the oxide, while the B-mode breakdown is considered to be extrinsic and defect-related. As the B-mode breakdown occurs near the operating electric field, it is very important to screen this mode in order to improve the reliability of the oxide film.

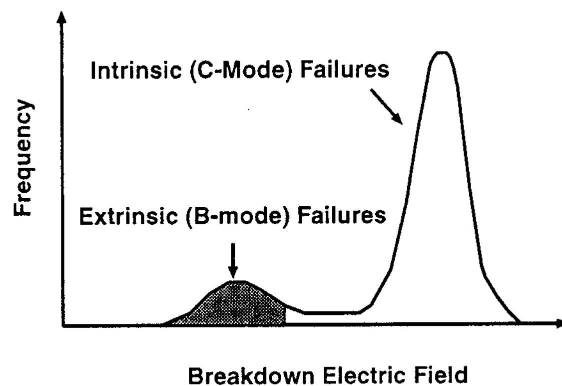

Fig. 1: Dependence on breakdown voltage of frequency of oxide film destruction.

Figure 2 is a conceptual illustration which shows the relation between the cumulative failure rate and operating time before and after screening. In this figure, to denote the life time required for the product, (a) shows the distribution before screening and (b) shows the distribution after screening without any damage recovery process. (b) indicates that the life time the oxide surviving from the screening is generally reduced due to the screening damage. The ideal screening, as (c) in the figure, should be such that all devices with life time less than t_0 is screened and the life time of survived devices is greater than t_0.

0-7803-4205-4/97/$10.00 ©1997 IEEE

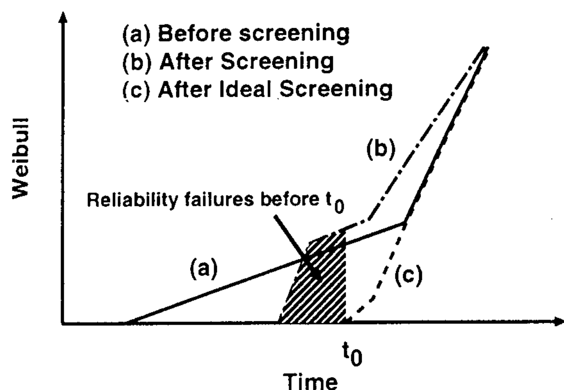

Fig. 2: The relation between the rate of failure caused by oxide film destruction and operating time, before and after screening.

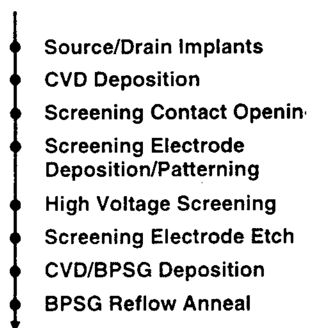

Fig. 3: Process flow chart of the in-process wafer-level screening inserted to the standard CMOS process.

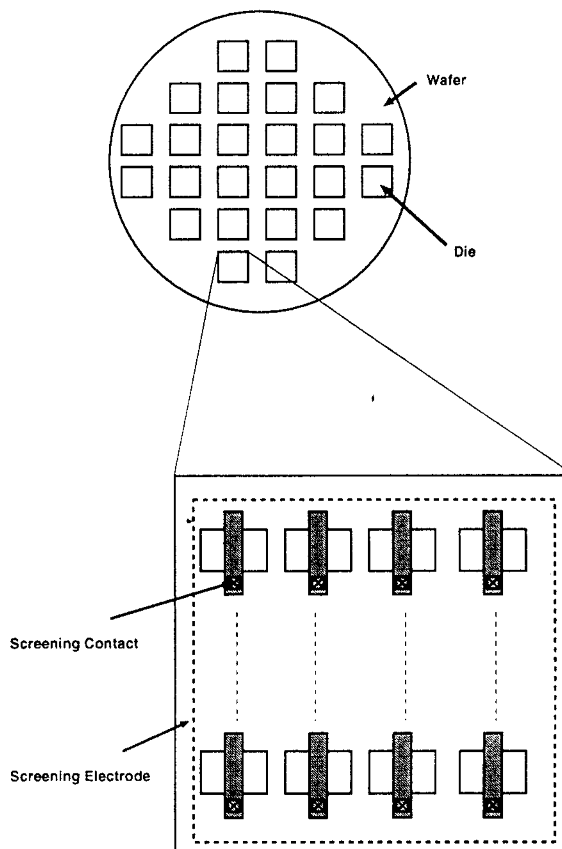

Fig. 4: Schematic view of a wafer and a die after the formation of the screening electrode.

However, the damage to the surviving oxide is unavoidable as long as the screening with stress is used. Therefore, some recovery process of the damage should be included in the ideal screening. Considering the well-known fact that the damage to oxide can be recovered by high-temperature annealing, it is possible to remove the damage if screening is performed before metal interconnection. In the following section, we describe the in-process wafer-level screening technique inserted to the standard CMOS technologies.

III. Detailed Description of the In-Process Wafer-Level Screening Method

A. *Process flow of the screening*

Figure 3 shows the process flow chart of the in-process wafer-level screening which is inserted to the standard CMOS process. After the source/ drain ion implantation is performed, a CVD film is deposited. This CVD film is used as a protective film which prevents the contamination by handling of wafer, etc., during the screening. Then contact holes are opened in order to connect the electrode for the screening to all gate poly Si of the transistors. After contact opening, poly Si is deposited on the whole area and patterned to form the screening electrodes which are connected to all gate electrodes simultaneously die by die. Figure 4 illustrates schematic view of a wafer and a die just after the patterning of the screening electrode. It is shown that each die has one screening electrode connecting all poly Si gates on the die in parallel.

At this stage, since metal interconnection has not been performed yet, the electrodes which can be used for applying voltage are the screening electrode and the back surface of the wafer. The well is connected to the backside of the wafer through the well-substrate junction. During screening stress, the voltage of the same polarity is applied to the gates of n-channel and p-channel transistors, and accordingly if p-channel transistors are in inversion, then n-channel transistors are in accumulation, and vice versa. As a result, sufficient carriers to the inversion channel area to form the inversion layer is necessary to apply expected stress voltage to the gate oxide for both transistors simultaneously. In order to supply the inversion carriers, we utilize light illumination over the wafer. By this method, it is possible to fully control the voltage applied to gate oxide. Figure 5 shows the schematic cross section of the device with the p-well/n-subtrate structure under screening with light illumination.

The position of the die which has defective oxides and is broken by the creening stress needs to be recorded for the later identification at the end of the line so that this die is not going into packaging process and shipping. After screening, the screening electrode is completely removed by chemical dry etching followed by CVD film deposition to fill up the screening contacts. Afterwards, the conventional BPSG film deposition and planalization reflow anneal are performed. This planalization reflow is usually done at about 800 °C or higher temperatures so that it can be used for the removal of the gate oxide damage due to the screening stress. The fabrication process after this is the same as that of the standard CMOS process.

For this screening technique, we need two additional masks compared to the standard CMOS process: one for screening contacts and one for screening electrode.

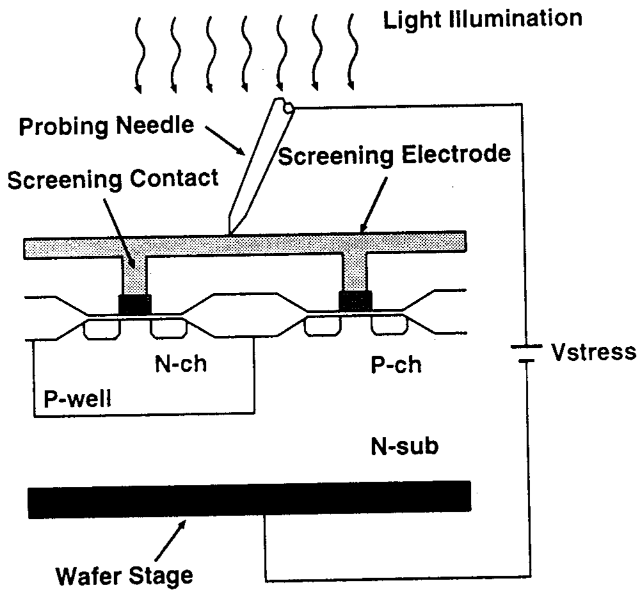

Fig. 5: Cross sectional view of a CMOS device under the in-process wafer-level screening with light illumination.

B. Screening voltage

As mentioned in the previous section, electron-hole pairs are created by light illumination to the wafer in order to apply the expected stress voltage simultaneously to both n-channel and p-channel gate oxides during the screening. In this section, we examine the electric field which is applied to the gate oxide of a CMOS device under screening.

Referring to the structure of Figure 4, we consider the case where the n-substrate is grounded and the voltage V_g is applied to the screening electrode. Note that in the actual screening, the current flowing in the system can be ignored because it is very small. If V_g is positive, for example, the p-channel transistor is in accumulation. Under light illumination, electron-hole pairs are generated in the channel region of the n-channel transistor. Generated electrons are used to form the inversion layer while generated holes are diffused in the p-well and eventually to the n-substrate through the p-well/n-substrate junction. Figure 6 shows the band diagrams for the CMOS structure for $V_g > 0$. Note that the system can be considered to be in equilibrium since the current is negligibly small for sufficiently small V_g.

Therefore the Fermi level is flat through the p-well and the n-substrate. From these band diagrams, the oxide voltage V_{ox}x can be expressed as

$$V_{ox} = V_g - V_{fb}(\text{p-well}) - \phi_s(\text{n-ch}) \qquad (V_g > 0) \qquad (1)$$

for the n-channel transistor and

$$V_{ox} = V_g - V_{fb}(\text{n-sub}) - \phi_s(\text{p-ch}) \qquad (V_g > 0) \qquad (2)$$

for the p-channel transistor, where V_{fb}(p-well) and V_{fb}(n-sub) are the flat band voltages for the p-well and the n-substrate, and ϕ_s(n-ch) and ϕ_s(p-ch) are the band bending of the surface potential of the n-channel and p-channel transistors, respectively.

On the other hand, if V_g is negative, the p-channel transistor is simply in inversion. For n-channel transistor, the p-well/n-substrate junction is reverse-biased so that some voltage drop V_j must be in this junction. Figure 7 shows the band diagrams for negative V_g. In this

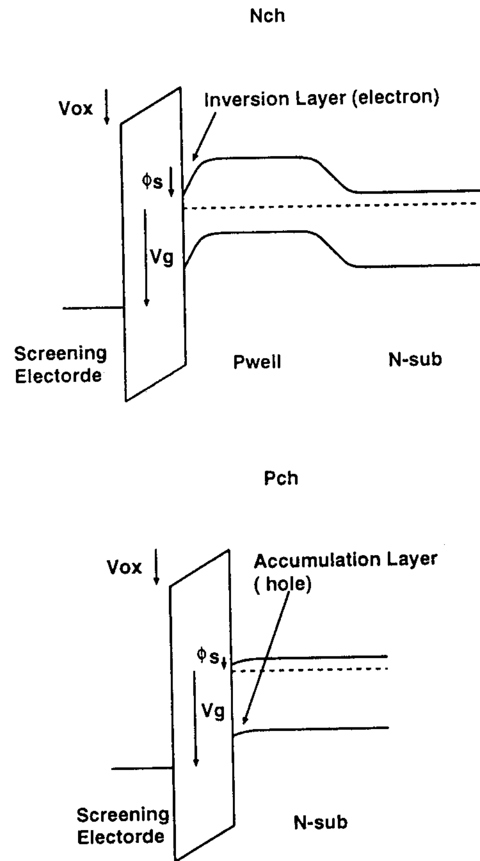

Fig. 6: Energy band diagrams of a p-well/n-substrate CMOS device for positive V_g.

case V_{ox} can be written as

$$V_{ox} = V_g - V_{fb}(\text{p-well}) - \phi_s(\text{n-ch}) - V_j \qquad (V_g < 0) \qquad (3)$$

for the n-channel transistor and

$$V_{ox} = V_g - V_{fb}(\text{n-sub}) - \phi_s(\text{p-ch}) \qquad (V_g < 0). \qquad (4)$$

for the p-channel transistor. In addition, from the charge neutrality condition, we have the following equation:

$$V_{ox}C_{ox} = V_j C_j(V_j) \qquad (V_g < 0), \qquad (5)$$

where $C_j(V_j)$ is the p-well/n-substrate junction capacitance, which depends on V_j.

C. The oxide field and temperature acceleration models for early TDDB failures.

The stress voltage, time and temperature must be determined adequately in order to screen the defective oxides and to minimize the damage to the non-defective oxides. In this section, dependence on voltage and temperature of early TDDB failures is investigated on the oxides of different thickness.

As the model for oxide field dependence of TDDB failures, two models, "E model" [8–10, 13] and "1/E model" [5–7, 12] have been proposed in the literature. As there is a significant difference in these two models in terms of the estimation of stress time in high oxide field which is equivalent to the stress time in low oxide field, it is a critical issue which model is used to estimate the stress time in the in-process

Table1: The oxide voltage in the p-well/n-substrate structure under the in-process wafer-lebel screening.

	p-ch	n-ch
$V_g > 0$	$V_g - V_{fb}$(n-sub) - ϕs(n-sub)	$V_g - V_{fb}$(p-well) - ϕs(p-well)
$V_g < 0$	$V_g - V_{fb}$(N-sub)-ϕs(n-sub)	$V_g - V_{fb}$(p-well) - ϕs(p-well) - V_j

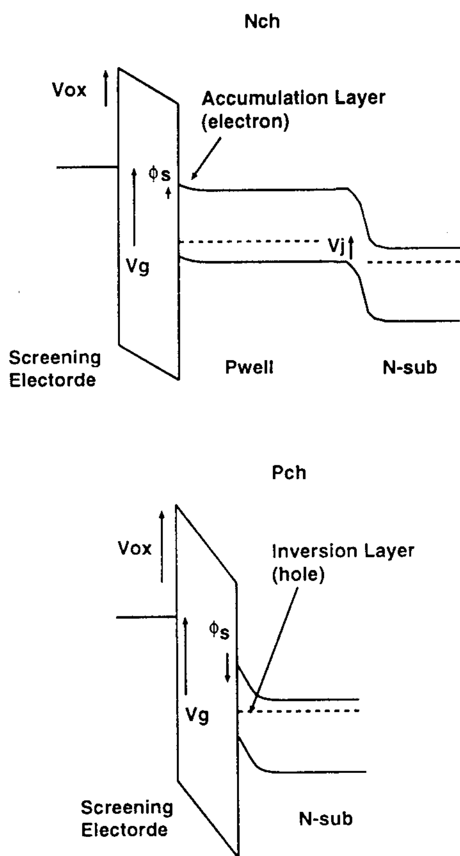

Fig. 7: Energy band diagrams of a p-well/n-substrate CMOS device for negative V_g.

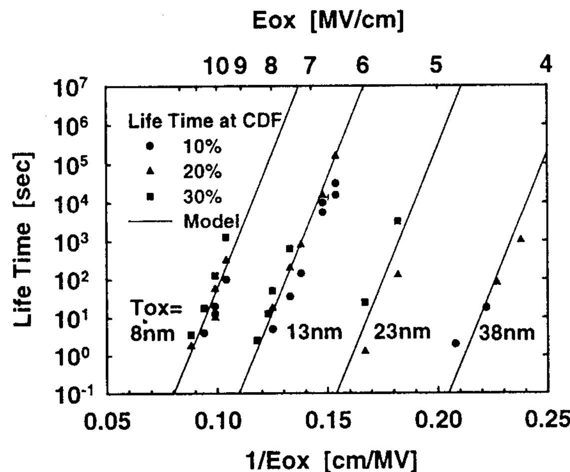

Fig. 8: TDDB life time at 10%, 20% and 30% cumulative failures as a function of reciprocal of oxide field for oxide thickness of $8nm$ to $38nm$. The life time is in proportion to the reciprocal of electric field ranging widely over the electric field. The straight line indicates $exp(320/Eox)$.

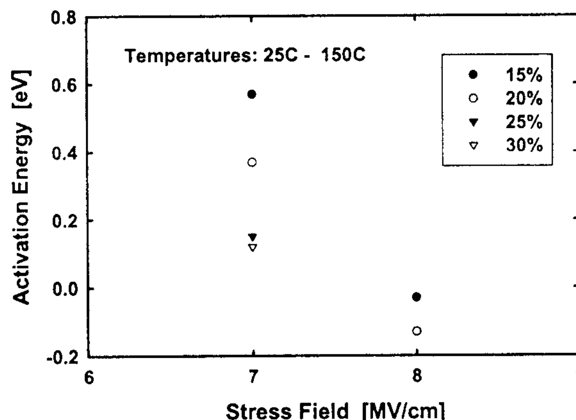

Fig. 9: The relation between the activation energy and stress electric field,calculated from TDDB life time at 15%, 20%, 25% and 30% cumulative failures of MOS capacitor with 15nm film thickness.

wafer-level screening. Then the relation between the oxide field and TDDB life time was examined, in order to decide the model to use. In Figure 8, TDDB life time of MOS capacitors with different oxide film thickness, with gate area is plotted as a function of the reciprocal of the oxide field. The data for the life time of early failures corresponding to 10%, 20% and 30% cumulative failures are shown in this plot. From this figure, it is clearly seen that the dependence of the early failures on the oxide field is well described with the 1/E model in the wide range of oxide field. The value of oxide field accelerating parameter G, which is defined as $exp(G/Eox)$, is fitted to be 320 from the data, and this is very consistent with the previously reported value [5,6].

The activation energy E_a is another parameter which accelerates the oxide failure. Figure 9 shows the dependence of activation energy on TDDB stress oxide field, for MOS capacitors with 15nm oxide. The life time is defined from 15% to 30% of cumulative failures. The data shows the activation energy increases as the oxide field decreases, and this behavior has been reported by McPherson and Baglee [8], and Boyko and Gerlach [11]. The value of activation energy which is shown here is quantitatively consistent with their reports. The data indicate the negative E_a values at 8 MV/cm. This might be unusual, but one possible explanation is as follows. According to the model that the oxide breakdown is due to the holes trapped in the oxide near the cathode, TDDB life time is determined by the amount of these holes. The rate of hole trapping decreases if de-trapping of trapped holes becomes significant. De-trapping is considered to be more significant at higher temperatures, which results in decreasing the effective rate of hole trapping and longer TDDB life time. This mechanism of oxide breakdown leads to the negative E_a values.

D. Recovery of the screening damage by high-temperature annealing.

As mentioned before, screening damage to non-defective oxides can be removed by high-temperature annealing after screening in this technique. In order to confirm this damage recovery, the failure

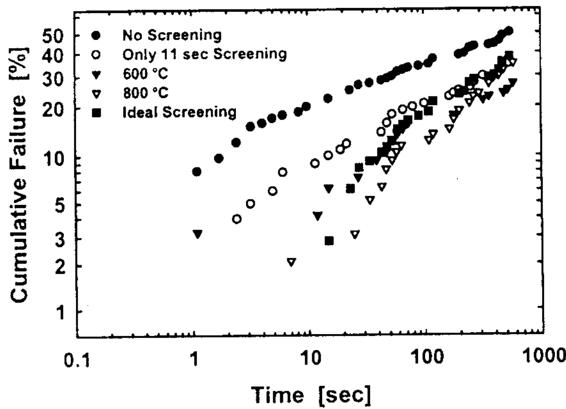

Fig. 10: TDDB results of capacitors subjected to the high voltage screening stress of 8 MV/cm for eleven sec followed by N_2 annealing at 600 °C and 800 °C. The results for capacitors with the screening but no annealing and control samples are also included.

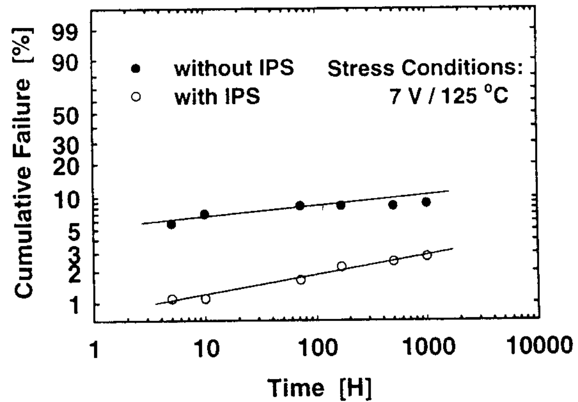

Fig. 11: Early failures due to the gate oxide TDDB as a function of time of dynamic life test (7 V and 125 °C) for 0.8μm CMOS LSIs with and without the in-process wafer-level screening.

distribution of stressed MOS capacitor samples was investigated as a function of annealing temperature. For this study, we used conventional capacitors with gate oxide thickness of 9.7 nm and gate area of 10 mm^2. The screening stress at 8 MV/cm for 11 sec at room temperature was performed on capacitors, which was followed by high-temperature annealing at 600 °C and 800 °C for 1 hour in N_2. Figure 10 plots cumulative TDDB failures as a function of stress time for capacitors with no stress, only stress and stress followed by 600 °C and 800 °C N_2 annealing. This figure also includes a plot for the failure distribution of ideal screening, which is calculated by subtracting the distribution just after the stress from that before the stress. This calculated ideal distribution clearly indicates that N_2 annealing at 600 °C or higher temperature is enough to recover the screening damage.

IV. Results of the In-Process Wafer-Level Screening for 0.8 μm CMOS LSIs

The in-process wafer-level screening technique described in detail in the previous section was applied to 0.8 μm CMOS logic LSI devices. The process technology used here has the p-well/n-substrate structure, 15 nm gate oxide and double-level Al metallization. The samples were split into two groups from the same lot: wafers with and without the in-process wafer-level screening. The screening stress field, temperature and time were 5.8 MV/cm, 25 °C and 5 sec, respectively. These values were determined such that the stress is approximately equivalent to the burn-in under 7 V, 10 hours and 125 °C, using the accelerating parameters in the previous section (G = 320 V/cm and $E_a = 0.6\ eV$). Figure 11 shows the Weibull plots of failures due to gate oxide TDDB during the dynamic life test of 7V and 125 °C for samples with and without the in-process wafer-level screening. In this figure, it is clearly shown that the early failures caused by oxide defects are significantly reduced by performing the in-process wafer-level screening. This result demonstrates the effectiveness of the in-process wafer-level screening which we have developed to reduce early TDDB failures.

V. Summary and Conclusion

We have developed a novel in-process wafer-level screening technique to eliminate infant mortality of LSI due to gate oxide defects. This technique makes it possible to stress all of gate oxides simultaneously at arbitrary high voltage for both n-channel and p-channel devices. It also shortens greatly the screening time and improves the reliability of oxide film. We have applied this technique to 0.8μm CMOS logic LSI and demonstrated the effectiveness of this technique for reducing early TDDB failures.

Q: Are you using this technique in the routine work?
A: No, we have not installed it in the routine work. This paper is a proposal.
Q: Do you have concerns about yield reduction when you use this technique?
A: No, we don't have any concerns about yield reduction because careful adjustment of the screening conditions (voltage, temperature and time) makes it possible to screen just defects which may eventually fail within the product life time.

Acknowledgments

The authors would like to acknowledge Katsuya Okumura of Microelectronics Laboratory and Kazuhiko Hashimoto of Semiconductor High Technologies Corp., for their helpful discussion and support for this study.

References

[1] S. Samata, M. Numano, T. Amai, Y. Matsushita, H. Kobayashi, A. Yamamoto, T. Kawaguchi, S. Nadahara, and K. Yamabe, *Ext. Abst. of the 1993 Electrochem. Soc. Fall Meeting*, 426 (1994).

[2] A. B. Joshi, R. Mann, L. Chung, M. Bhat, T. H. Cho, B. W. Min, and D. L. Kwong, "Suppressed Process-Induced Damage in N_2O-annealed SiO_2 Gate Dielectrics,"*Proc. 1995 IRPS*, 156 (1995).

[3] T. Ohmi, K. Nakamura, and K. Makihara, "Highly-Reliable Ultra-Thin Oxide Formation Using Hydrogen-Radical-Balanced Steam Oxidation Technology," *Proc. 1994 IRPS*, 161 (1994).

[4] J. C. King, W. Y. Chan and C. Hu,"Efficient Gate Oxide Defect Screen for VLSI Reliability," *1994 IEDM Tech. Digests*, 597 (1994).

[5] I. C. Chen, S. Holland, and C. Hu, "Electrical Breakdown in Thin Gate and Tunneling Oxides," *IEEE Trans. Electron Devices*, **ED-32**, 413 (1985).

[6] I. C. Chen, and C. Hu, "Accelerated Testing of Time-Dependent Breakdown of SiO_2,"*IEEE Electron Device Lett.*, **EDL-8**, 140 (1987).

[7] J. Lee, I. C. Chen and C. Hu, "Modeling and Characterization of Gate Oxide Reliability," *IEEE Trans. Electron Devices*, **ED-35**, 2268, (1988).

[8] J. W. McPherson and D. A. Baglee, "Acceleration Factors for Thin Gate Oxide Stressing," *Proc. 1985 IRPS*, 1 (1985).

[9] D. Crook, "Method of Determining Reliability Screens for Time Dependent Dielectric Breakdown," *Proc. 1979 IRPS*, 1 (1979).