Fig. 5. Effectiveness of the burn-in simulation depends on $\Gamma$. The results shown are for circuit c880 in the ISCAS '85 benchmark circuit set. The number by each symbol is the number of vectors in the input vector sequence used during the burn-in simulation. $X_{\text{ref}}$ was set at 8 nm for the calculation of $\Gamma$.

Fig. 6. Convergence characteristics of TDDB failure predictions.

25 input vectors. The rapid convergence is due to the fact that the circuit failure, to the first-order, is based on the sum of stress times.

V. CONCLUSION

We have developed a fast oxide reliability module for digital CMOS circuits for the circuit-level reliability simulator BERT. The stress activity on the gate oxides is analyzed using a logic simulator. The burn-in simulation predicts a significant improvement to the circuit reliability after a burn-in pass. The use of the figure-of-merit $\Gamma$ is a convenient way to gauge the effectiveness of the burn-in simulation without a full two-pass simulation.

REFERENCES


Relaxation of Acceptance Limits (RAL):
A Global Approach for Parametric Yield Control of
0.1-$\mu$m Deep Submicron MOSFET Devices

Renate Sitte, Sima Dimitrijev, and H. Barry Harrison

Abstract—An alternative method to fixed quality acceptance limits for in-line yield control is proposed. Our study is based on a sensitivity analysis, which has revealed that conventional parametric yield-control techniques using fixed in-line acceptance (tolerance) limits, as traditionally used in semiconductor manufacturing, are not efficient in deep submicron-size devices.

I. INTRODUCTION

Improvements in integrated circuit (IC) manufacturing techniques and equipment have been dictated mainly by the need for changes and refinement of semiconductor technology itself. As soon as technology catches up with the demand of refinement, the frontiers are pushed further forward. A factory, initially equipped with state of the art equipment to guarantee a well-controlled production environment, is soon pushed by competition and market demands into extreme production, stretching conditions to the limitations of equipment. With conditions close to the limitations of equipment there is little slack, and processes are less controllable. For example, a replication equipment may be well controlled for 0.5-$\mu$m, but not for 0.25-$\mu$m line widths. With current equipment resolution, the main limitation of conventional process control is the inability to control fluctuations to the level required in deep submicron devices. This is because with the downscaling, some physical effects may become more pronounced and dominant over others, changing the effect of manufacturing fluctuations on the device. Several 0.1-$\mu$m MOSFET’s have been

Manuscript received December 9, 1993; revised December 24, 1994.
R. Sitte is with the Defence Science & Technology Organization, Information Technology Division, Salisbury, South Australia.
S. Dimitrijev and H. Harrison are with the School of Microelectronic Engineering, Faculty of Science and Technology, Griffith University, Nathan, Queensland, Australia.
IEEE Log Number 9412873.
IEEE TRANSACTIONS ON SEMICONDUCTOR MANUFACTURING, VOL. 8, NO. 3, AUGUST 1995

commercial production at this level may come to fruition in the future. While the published results claim improved device electrical characteristics, these device dimensions raise the important practical question as to whether it will be possible to reproduce these devices in larger numbers with acceptable yields.

It is the purpose of this paper to present an efficient method for parametric yield control. This method is based on a novel in-line quality acceptance criterion where the accept/reject decision is delayed until further on in the process, because the in-line measurement information can be used in a more profitable way. It is also shown in this paper why traditional fixed-quality acceptance (tolerance) limits for in-line measurements are unsuitable to achieve high parametric yield for deep submicron devices.

II. BOUNDARIES OF THE CRITICAL PROCESSING PARAMETERS

To gain insight into the effects of process parameters on deep submicron devices a study [5], [6] based on simulation using MINIMOS [7] has been carried out for a 0.1-μm deep submicron MOSFET designed at IBM [1], using fluctuations which can be found typically in modern integrated circuit manufacturing. The suitability of MINIMOS as a simulation tool for deep submicron devices has been asserted previously [8]. In that study attention has been focused on four main device parameters: the threshold voltage, the transconductance, the drain current when the transistor is off, and the substrate current.

The study revealed that only a few processing parameters are critical contributors to the device parameter manufacturing fluctuations. In particular, for the threshold voltage, the Pareto Analysis [9] results are as follows: that the gate oxide thickness (TOX) would contribute typically with more than 60% to the threshold voltage fluctuation, the gate length (LENG) would add another 20%, and the energy for threshold voltage adjustment implants (ENER) would add further 15% to the overall fluctuation. Thus, it is on these three processing parameters on which control should be exerted.

A set of threshold voltage values were obtained by simulation, using different combinations of the three critical processing parameter values, taken at regularly spaced points laying within two standard deviations around the processing specification data (recipe). From this set, all those data points were selected, which would fall into the acceptable range for the threshold voltage. We chose the range of 0.1 V ≤ Vth ≤ 0.25 V as acceptable. Within this selected subset, the largest and smallest occurring values for each of the critical processing parameters were taken. This corresponds to finding upper and lower boundaries for in-line measurements, or tolerance limits, which would make an acceptable threshold voltage possible. Beyond these extreme boundaries no acceptable threshold voltage can be found utilizing any combination. The extreme possible values are listed in Table I.

III. FIXED-QUALITY ACCEPTANCE LIMITS: A LOW YIELD SOLUTION

The contour map provides a graphical illustration that fixed tolerance limits of in-line measurements are difficult to determine. Consider a range of values centered around the specified processing values (recipe), for example between 3.5 and 5.5 nm for TOX, and 0.09 and 0.11 μm for LENG. A fixed-limit acceptance region (tolerance region) would be the rectangular shape formed between intersecting pairs of upper and lower lines, set anywhere between the dotted and dashed lines.

It has been found, however, that not necessarily all combinations of processing parameter values occurring within these extreme processing step boundaries lead to acceptable threshold voltages. This is because the geometric space of values for which the threshold voltage is acceptable is not regularly shaped. This can be seen in Fig. 1, from the contour map of the threshold voltage as a function of gate oxide thickness and gate length. The triplets of contours correspond to energies of threshold voltage adjustment implants of 14.7, 15, and 15.4 keV, respectively. The shaded area corresponds to the acceptance region for threshold voltage and RAL. A fixed-limit acceptance region (tolerance region) would be the rectangular shape formed between intersecting pairs of upper and lower lines, set anywhere between the dotted and dashed lines.

![Fig. 1. Contour map of the threshold voltage as a function of gate oxide thickness and gate length. The triplets of contours correspond to energies of threshold voltage adjustment implants of 14.7, 15, and 15.4 keV, respectively. The shaded area corresponds to the acceptance region for threshold voltage and RAL. A fixed-limit acceptance region (tolerance region) would be the rectangular shape formed between intersecting pairs of upper and lower lines, set anywhere between the dotted and dashed lines.](image)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Lower limit</th>
<th>Upper limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOX (nm)</td>
<td>3.5</td>
<td>5.5</td>
</tr>
<tr>
<td>LENG (μm)</td>
<td>0.09</td>
<td>0.11</td>
</tr>
</tbody>
</table>

It should be noted that the acceptance region is only a shaded area. A fixed-limit acceptance region would be a more narrow rectangular area, which also are intersecting with the shaded area, are workable MOSFET's. Other sets of fixed limits can be chosen anywhere between the dotted and dashed lines in Fig. 1, and the acceptance region would be the rectangular shaped area between intersecting pairs of upper and lower lines.
but process them through the whole line, which is an expensive exercise. Too narrow boundaries would exclude too many potentially good wafers. The solution is the relaxation of acceptance limits RAL.

IV. RELAXATION OF ACCEPTANCE LIMITS: THE HIGH YIELD ALTERNATIVE

The RAL method is a quality-acceptance criterion, based on the concept that several combinations can give the effect of the same result. A device parameter may be found to be within acceptable tolerance limits, although individual in-line measurements appear to be beyond tolerance, or vice versa. The technique consists in relaxing the quality acceptance limits (tolerance limits) to be anywhere within the acceptance region for the threshold voltage, i.e., the shaded area in Fig. 1. From the results of in-line measurements, the accept/reject criterion is flexibly found. By monitoring the progress after each step, it can be decided whether the wafer will be processed further, or should be removed from the line. This can be easily done by either solving a second order polynomial, or by consulting the contour map. The process is described in more detail below for the case of threshold voltage.

A. Application of the RAL Criterion

The expression of the second order polynomial adapted to our case is

\[
V_{th} = a_0 + a_1(\text{TOX}) + a_2(\text{TOX})^2 + a_3(\text{ENER}) + a_4(\text{ENER})^2 + a_5(\text{LENG}) + a_6(\text{LENG})^2 + a_7(\text{TOX})(\text{ENER}) + a_8(\text{TOX})(\text{LENG}) + a_9(\text{ENER})(\text{LENG})
\]

(1)

where \(a_0, a_1, \ldots, a_9\) are the polynomial coefficients. The data needed to determine the polynomial coefficients \(a_0, a_1, \ldots, a_9\) by a fitting procedure were generated by simulating the threshold voltage, in all possible combinations of values of the critical processing parameters involved, as described in Section II.

This polynomial, which is a mathematical model approximating the threshold voltage as a function of TOX, ENER, and LENG, can be used throughout the processing of a wafer to indicate whether a wafer is likely to produce devices with acceptable threshold voltage. Initially the variables of the polynomial are set to the processing specification value of the critical steps. By subsequently replacing these values with the values of in-line measurements of the steps already processed, and solving the polynomial, the expected threshold voltage is obtained. The pattern to be used is \(V_{th} = f(\text{TOX}, \text{ENER}, \text{LENG})\) where \(f\) is the polynomial function as indicated in (1). For example, the polynomial expression \(V_{th} = f(4.8, 15.3, \text{and} 0.104)\) indicates an expected threshold voltage of 0.2695 V, and the wafer would be discarded erroneously. LTL and UTL are the lower and upper tolerance limits, respectively.

Any choice of too wide boundaries would include too many wafers for which an acceptable threshold voltage cannot be expected, and vice versa, wafers falling into the gray shaded area but outside the intersection of striped areas, would be discarded erroneously. LTL and UTL are the lower and upper tolerance limits, respectively.

Fig. 2. Contour map of the threshold voltage as a function of gate-oxide thickness and gate length and energies of threshold voltage adjustment implants. Examples of fixed-limit acceptance region: (a) centered around specified processing values (recipe); (b) wide range (extreme possible boundaries); (c) minimum range (all overlapping boxes). In all cases the acceptance region would be the intersection of the striped areas. Wafers within that intersection, but not overlapping with the gray shaded area, would erroneously be accepted, and vice versa, wafers falling into the gray area but outside the intersection of striped areas, would be discarded erroneously. LTL and UTL are the lower and upper tolerance limits, respectively.
B. RAL Simulation Experiment

A Monte Carlo simulation experiment was set up to compare the yield benefits of RAL with fixed-limit sampling. This experiment consisted in generating 500 data sets of processing parameters, simulating random fluctuations around the processing specification value (recipe). The data sets were fed into a device simulator [7] to determine the threshold voltage. The data sets or samples, were then subject to “quality inspections” following the same sequence as they would be for in-line measurements, applying the criteria of fixed-tolerance limits of the overlapping box. Samples failing the acceptance criteria were subsequently excluded. The same original sample sets were also examined with criteria or fixed-tolerance limits. This experiment establishes superiority of RAL as a quality assessment procedure. Further research will be directed towards refinements and combining benefits into one integrated quality-assurance program.

V. CONCLUSION

In this paper RAL has been presented as a novel technique for processing step in-line quality assessments for deep submicron MOSFET’s. The method is opposed to traditionally fixed in-line quality acceptance limits (tolerance limits). It has been shown that for future 0.1-μm technology a quality assessment using fixed-tolerance limits is not recommendable. This is because the geometric shape of the threshold voltage acceptance region is irregular, making an unambiguous correspondence to fixed-tolerance limits impractical. With RAL a yield increase of almost 30% has been found. This establishes superiority of RAL as a quality assessment procedure. Further research will be directed towards refinements and combining benefits into one integrated quality-assurance program.

REFERENCES