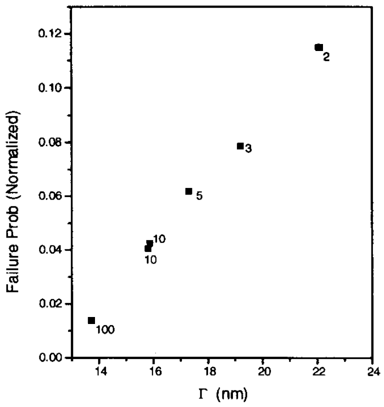

Fig. 5. Effectiveness of the burn-in simulation depends on Γ. The results shown are for circuit c880 in the ISCAS '85 benchmark circuit set. The number by each symbol is the number of vectors in the input vector sequence used during the burn-in simulation. $X'_{\rm eff}$ was set at 8 nm for the calculation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

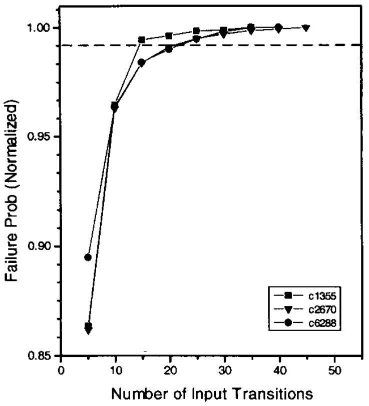

Fig. 6. Convergence characteristics of TDDB failure predictions.

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-demerit Γ is a convenient way to gauge the effectiveness of the burn-in simulation without a full two-pass simulation.

References

[1] A. Berman, "Time-zero dielectric reliability test by a ramp method," in *IEEE Int. Reli. Phys. Symp.*, 1981, pp. 204–208.

[2] I. C. Chen, S. E. Holland, and C. Hu, "A quantitative physical model for time-dependent breakdown in SiO2," in *IEEE Int. Reli. Phys. Symp.*, 1985, pp. 24–29.

[3] E. Rosenbaum, P. M. Lee, R. Moazzami, P. K. Ko, and C. Hu, "Circuit reliability simulator—Oxide breakdown module," in *IEEE Int. Electron Dev. Meeting*, 1989, pp. 331–334.

[4] J. C. Lee, I. C. Chen, and C. Hu, "Modeling and characterization of gate oxide reliability," *IEEE Trans. Electron Dev.*, vol. ED-35, pp. 2268–2278, Dec. 1988.

[5] R. H. Tu *et al.*, "BERT—Berkeley reliability tools," Electrical Research Laboratory Technical Memorandum M91 107, University of California at Berkeley, Dec. 1991.

[6] E. Rosenbaum and C. Hu, "High-frequency time-dependent breakdown of SiO2," *IEEE Trans. Electron Dev. Lett.*, vol. 12, pp. 267–269, June 1991.

[7] R. Moazzami and C. Hu, "Projecting gate oxide reliability and optimizing reliability screens," *IEEE Trans. Electron Dev.*, vol. 37, pp. 1643–1657, July 1990.

[8] F. Brglez, P. Pownall, and R. Hulm, "Recent algorithms for gate-level ATPG with fault simulation and their performance assessment," in *IEEE Int. Symp. Circuits Syst.*, 1985, pp. 664–698.

[9] F. Brglez, D. Bryan, and K. Kozminski, "Combinational profiles of sequential benchmark circuits," in *IEEE Int. Symp. Circuits Syst.*, 1989, pp. 1929–1934.

Relaxation of Acceptance Limits (RAL): A Global Approach for Parametric Yield Control of 0.1-μm 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-μm, but not for 0.25-μm 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-μm 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, Australia.

IEEE Log Number 9412873.

TABLE I
BOUNDARIES FOR THE MOST INFLUENTIAL PROCESSING PARAMETERS, FOR WHICH ACCEPTABLE THRESHOLD VOLTAGE VALUES CAN BE FOUND

	TOX		ENER		LEN	
Unit	nm		keV		μm	
Processing specification	4.5		15		0.1	
Std. dev (σ)	1.5		0.15		0.06	
	lower	*upper*	*lower*	*upper*	*lower*	*upper*
Extreme possible limits	3*	6.6	14.7	15.4	0.08	0.128
Overlapping box limits	3.3	5	14.7	15.4	0.084	0.105

* lower physical limit

designed and manufactured [1]–[4], providing reason to expect that 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 implant (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 $0.1\text{ V} \leq V_{\text{th}} \leq 0.25\text{ V}$ 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.

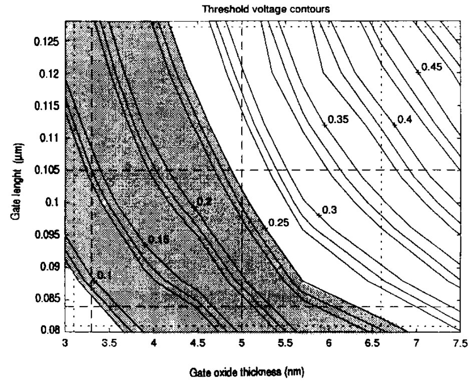

Fig. 1. Contour map of the threshold voltage as a function of gate oxide thickness and gate length. The triplets of contours correspond to energies of threshold voltage adjustment implants of 14.7, 15, and 15.4 keV, respectively. The shaded area corresponds to the acceptance region for threshold voltage 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.

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 TOX and LENG, and ENER; the shaded area indicates the acceptance region for the threshold voltage—i.e., what *should* be accepted. For example, the combination of TOX = 4.8 nm, ENER = 15.3 keV, and LENG = 0.112 μm produces an unacceptable threshold voltage of 0.266 V, while an apparently worse combination of TOX = 6.60 nm, ENER = 15.4 keV, and LENG = 0.08 μm produces an acceptable threshold voltage of 0.236 V. On this graph, the extreme upper and lower possible values for TOX and LENG are indicated as dotted lines, and the triplets of contours correspond to energies of threshold voltage adjustment implants of 14.7, 15, and 15.4 keV, respectively.

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. This choice of tolerance limits would lead to the acceptance of all MOSFET's with values laying on the intersection of these two sets of values. It is indicated by the striped area shown in Fig. 2(a). However only those MOSFET's in the striped 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. If the extreme possible values were used (intersection of the striped areas in Fig. 2(b)), almost half of the wafers produced would be discarded at the end of processing, because they would be in the upper right area of the rectangle, i.e., the threshold voltage is greater than 0.25 V. If acceptance region is narrowed to the area where an acceptable threshold voltage is *always* obtained ("overlapping box" on Table I, and intersection of the striped regions on Fig. 2(c)), the yield would be reduced to be almost half of its potential.

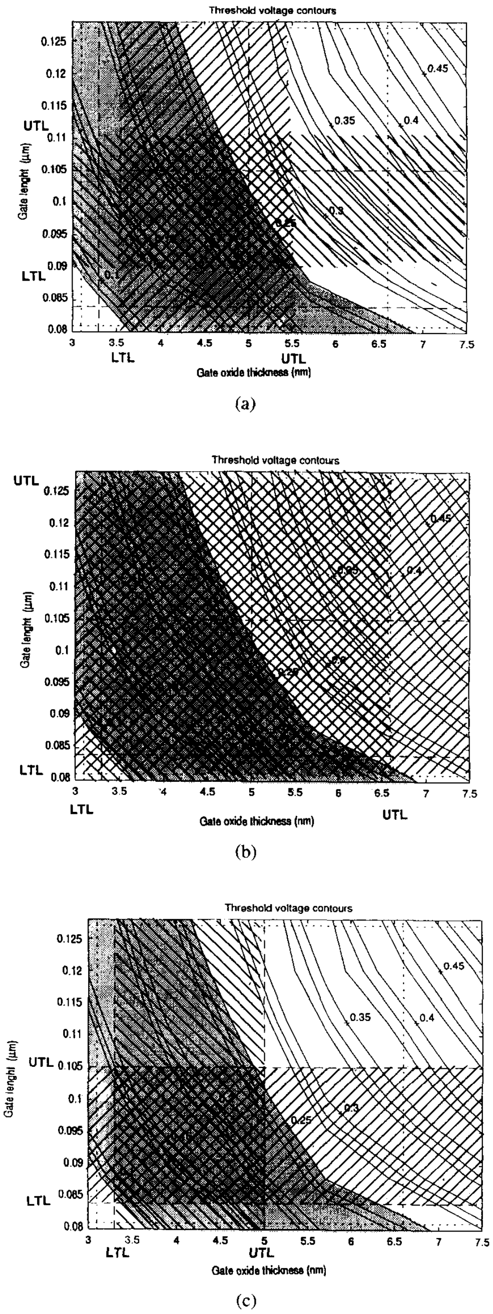

Fig. 2. Contour map of the threshold voltage as a function of gate-oxide thickness and gate length and energies of threshold voltage adjustment implants. Examples of fixed-limit acceptance region: (a) centered around specified processing values (recipe); (b) wide range (extreme possible boundaries); (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.

Any choice of too wide boundaries would include too many wafers for which an acceptable threshold voltage cannot be expected, 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

$$\begin{aligned} V_{\rm th} = {} & a_0 + a_1(\mathrm{TOX}) + a_2(\mathrm{TOX})^2 + a_3(\mathrm{ENER}) \\ & + a_4(\mathrm{ENER})^2 + a_5(\mathrm{LENG}) + a_6(\mathrm{LENG})^2 \\ & + a_7(\mathrm{TOX})(\mathrm{ENER}) + a_8(\mathrm{TOX})(\mathrm{LENG}) \\ & + a_9(\mathrm{ENER})(\mathrm{LENG}) \end{aligned} \tag{1}$$

where $a_0, a_1, \ldots a_9$ are the polynomial coefficients. The data needed to determine the polynomial coefficients $(a_1, a_2, \ldots a_9)$ 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 $\mathrm{V_{th}} = f$ (TOX, ENER, and LENG) where f is the polynomial function as indicated in (1). For example, the polynomial expression $\mathrm{V_{th}} = f$ (4.8, 15.3, and 0.112) indicates an expected threshold voltage of 0.2695 V, which is unacceptable and the wafer has to be removed from further processing. If the gate length had been 0.104 μm, the evaluation of the polynomial $\mathrm{V_{th}} = f$ (4.8, 15.3, and 0.104) would have produced an expected threshold voltage of 0.2458 V, and the wafer would be processed further.

For practical purposes the second order polynomial fit needs to be done only once for a MOSFET design. To avoid numerical problems caused by the magnitude of the quantities involved, all numbers have to be normalized with respect to their processing specification value, before being used in polynomial calculations. After the polynomial evaluations, the normalization has to be reverted back to the actual values of the threshold voltage.

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 RAL criteria. A final comparison of both sets revealed that only 27.7% would be accepted following a fixed tolerance limits procedure, 56.9% would be accepted with RAL, i.e., 29.2% more acceptable wafers were found by the RAL procedure than with the fixed tolerance limits, doubling the yield. The yield increase is due to recognizing potentially good wafers. Further comparisons showed good accuracy between the threshold voltages found by polynomial approximation and the Monte Carlo simulations. One can conclude that the polynomial approximation is a good estimate for threshold voltage predictions.

It should be noted that RAL cannot guarantee that an acceptable threshold voltage will be achieved for that wafer. Nevertheless, the wafers kept in the production line have a high likelihood of producing acceptable threshold voltages. This means that more wafers are processed until the end, which might have been otherwise discarded by unrealistic fixed-tolerance limits. On the other hand, not all wafers kept in production up to and including the implant for threshold voltage adjustment will stay in the line. It is most likely after the third critical processing step, i.e., the etching of the gate length, when a clear cut decision is taken. The saving is in avoiding further processing wafers that are unlikely to produce acceptable threshold voltage. RAL is suitable for relatively small deviations from recipe, when a process is well tuned. For larger deviations, for example when the gate oxide is thicker than the upper extreme limit, a forward correction technique such as dynamic design processing (DDP) can be used [8].

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

[1] G. A. Sai-Halasz, M. R. Wordeman, D. P. Kern, E. Ganin, S. Rishton, D. S. Zichermann, H. Schmid, M. R. Polcari, H. Y. Ng, P. J. Restle, T. H. P. Chang, and R. B. Dennard, "Design and experimental technology for 0.1-μm gate-length low temperature operation MOSFET's," *IEEE Trans. Electron Dev. Lett.*, vol. EDL-8, pp. 463–466, 1987.

[2] M. Aoki, T. Ishii, T. Yoshimura, Y. Kiyota, S. Iijima, T. Yamanaka, T. Kure, K. Ohju, T. Nishida, S. Okazaki, K. Seki, and K. Shimohigashi, "Design and performance of 0.1-μm CMOS devices using low impurity channel transistors (LICT's)," *IEEE Trans. Electron Dev. Lett.*, vol. 13, no. 1, pp. 50–52, Jan. 1992.

[3] M. Iwase, T. Mizuno, M. Takahashi, H. Niiyama, M. Fukumoto, K. Ishida, S. Inaba, Y. Takigami, A. Sanda, A. Toriumi, and M. Yoshimi, "High performance 0.1-μm CMOS devices operating at room temperature," *IEEE Trans. Electron Dev. Lett.*, vol. 14, no. 2, pp. 51–53, Feb. 1993.

[4] Y. Taur, S. Cohen, S. Wind, T. Lii, C. Hsu, D. Quinlan, C. A. Chang, D. Buchanan, P. Agnello, Y. Mii, C. Reeves, A. Acovoc, and V. Kesan, "Experimental 0.1-μm p-channel MOSFET with p^+-polysilicon gate on 35-Å gate oxide," *IEEE Trans. Electron Dev. Lett.*, vol. 14, no. 6, June 1993.

[5] R. Sitte, S. Dimitrijev, and H. B. Harrison, "Sensitivity of 0.1-μm MOSFET's to manufacturing fluctuations," *Electron. Lett.*, vol. 29, no. 15, pp. 1345–1346, July 1993.

[6] ——, "Device parameter changes caused by manufacturing fluctuations of deep submicron MOSFET's," *IEEE Trans. Electron Dev.*, vol. 42, no. 11, pp. 2210–2215, 1994.

[7] "MINIMOS," Technische Universität Wien, Institut für Mikroelektronik, Austria, 1990.

[8] R. Sitte, S. Dimitrijev and H. B. Harrison, "The effect of dynamic design processing for yield enhancement in the fabrication of deep-submicron MOSFET's," *IEEE Trans. Semiconduct. Manufact.*, vol. 7, no. 1, pp. 92–96, Feb. 1994.

[9] J. Banks, *Principles of Quality Control.* New York: Wiley, 1989.

[10] R. Sitte, S. Dimitrijev, and H. B. Harrison "Methods for parametric yield control for future 0.1-μm deep submicron MOSFET manufacturing," in *Manufacturing Process Control for Microelectronic Devices and Circuits* (Proceedings SPIE 2336), A. G. Sabnis, Ed. Bellingham: SPIE, 1994, pp. 202–207.