Statistically Independent Soft Breakdowns Redefine Oxide Reliability Specifications

M. A. Alam, R. K. Smith, B. E. Weir, and P. J. Silverman
Agere Systems, 600 Mountain Avenue, Murray Hill, NJ 07974, alam@agere.com

Abstract

The statistics of soft-breakdown events in ultrathin oxide transistors are studied. By using new theoretical techniques and quantitative analysis, we demonstrate that spatial and temporal correlations among the successive breakdown events are weak. This allows us to redefine the standard specification of oxide reliability and suggest that ultrathin oxides will be far more reliable and fault-tolerant than has been assumed thus far.

Introduction

The standard definition of oxide reliability requires fewer than 100 parts per million IC failure (defined as a soft breakdown [SBD] for even one transistor in an IC) within 10 years. By this definition, n-MOS remains reliable down to at least 1.1-1.3 nm oxide thickness [1], but p-MOS may not [2] (primarily due to minority ionization[3]). However, since the circuits do continue to operate after the first SBD [4-7], it has been suggested that the standard reliability specification is too restrictive, and should be redefined. In this paper, we establish the theoretical framework for such a redefinition with a systematic and thorough statistical analysis of the degradation phenomena after the first SBD event [see Fig. 1a for details]. Our analysis demonstrates that consecutive SBDs in a given transistor are essentially uncorrelated, allowing us to accurately predict the progression of oxide degradation and show how area- and percentile- scaling should be redefined under these circumstances.

Statistical distribution of soft-breakdown events

ICs may continue to operate immediately after a SBD [4-7], but if the localized current through the first breakdown spot forces a subsequent SBD (Fig. 1b) to occur in close proximity (*space-correlation*) or in quick succession (*time-correlation*), then the transistor could rapidly become unstable, and adoption of soft-broken oxides would not increase the functional lifetime of the ICs significantly. Therefore, it is critically important to precisely establish the degree of correlation among the positions and times of breakdown events. This we do by first developing a theoretical model for uncorrelated breakdown suitable for easy comparison to the measured data. Imagine dividing the dielectric thin film, with area A_{film} and thickness T_{film} into a 3D cubic lattice. The number of defective cells, assigned randomly in the 3D lattice to mimic uniform defect generation, increases monotonically with time. A "short" forms when all cells of a vertical column become defective, creating a low resistance path between the electrodes. This finite-size percolation model [6] is easy to analyze and its results (to be discussed later) will help establish the degree of correlation among breakdown events.

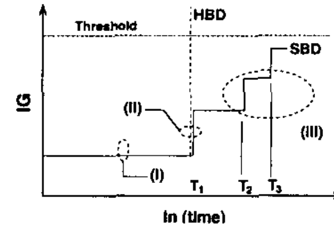

Fig. 1a: Oxide degradation can be divided into three different regimes: (I) the build-up randomly generated defects up to the first BD time T_1 [1], (II) the power dissipation-dependent transition to SBD or HBD [6], and (III) subsequent SBD or HBD events, $T_2, T_3, \ldots$). In this paper, we focus on a systematic analysis of the third stage of oxide degradation.

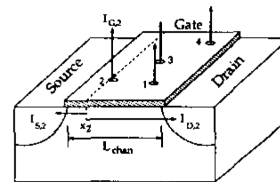

Fig. 1b: Separate soft breakdown events are drawn schematically. As shown, current through a breakdown spot is balanced by currents from the source and the drain. The ratio of the source to drain currents can be used to locate the position of the breakdown spot (*e.g.* x_2 as shown above) within the channel L_{chan}.

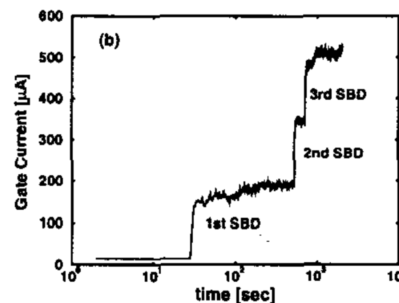

Fig. 1c: A representative measurement of gate current from which time to successive breakdowns and location of breakdown spots were determined. The data is obtained from a NFET stressed in inversion at 3.5 volts, with oxide thickness of 1.6 nm, and gate area of 2.5 μm^2.

To check the *space-correlation* among the breakdown spots, we determined (see Fig. 1c and [8]) positions of consecutive soft breakdowns of ~ 100 transistors (Fig. 2a).

0-7803-7462-2/02/$17.00 ©2002 IEEE

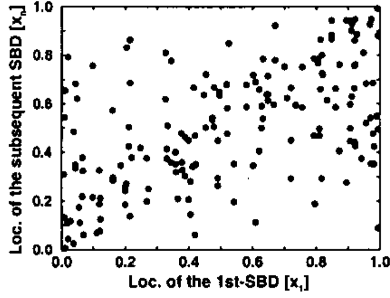

Fig. 2(a): The measured soft breakdown spots of $\sim$ 100 samples are shown. If the first breakdown localized the subsequent breakdown events, the space-correlation would have been strong, and the datapoints would have clustered along the diagonal. Measured data does not indicate such strong spatial correlation.

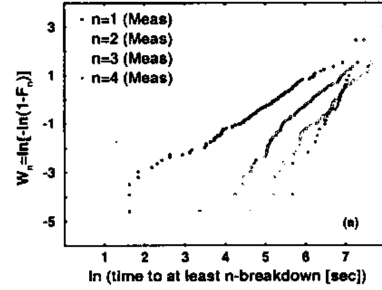

Fig. 3(a): Weibull plots for devices with at least n breakdown events, W_n. The 1.6 nm oxides with gate area of 2.5 μm^2 were stressed in inversion at 3.5 volts.

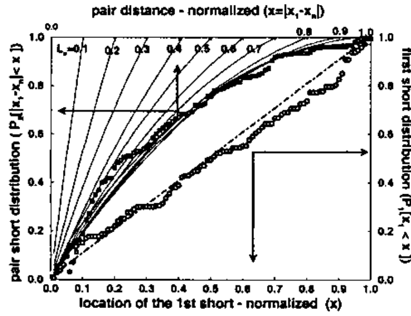

Fig. 2(b): The PDF of the distance between pairs of breakdown spots for each device, given by the difference of the x- and y-coordinates of each point in Fig. 2a, is plotted. The thin lines are theoretical plots with various degrees of spatial correlation present (as marked), with $L_c \rightarrow 0$ for completely correlated breakdown, and with $L_c \rightarrow 1$ for completely uncorrelated ones. The measured data are well represented by a correlation distance of 0.83, indicating weak correlation.

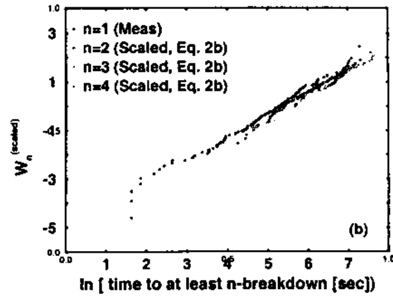

Fig. 3(b): When the Weibull slopes of Fig. 3a are scaled using Eq. 2b, they coincide, indicating very weak *time-correlation*. More detailed estimate puts the correlation upper-bound to be less than 10 %.

The datapoints do not cluster along the diagonal, indicating that the second SBD events do not necessarily occur in close proximity to the first SBD. Therefore, the spatial-correlation must be weak. For a more quantitative analysis using the finite-size percolation model, note that if the defect generation is spatially random, the location of the first vertical short is also random: the cumulative probability distribution function (CDF) of finding the first short at any location x_i (appropriately normalized) on the film is

$$P_1\left[x_i < x\right] = \alpha x + \beta, \qquad \text{(1-a)}$$

with $\alpha = 1$, and $\beta = 0$. If the presence of one short at x_i influences subsequent ones at x_j to occur within a correlation distance of L_c, then the pair CDF must follow

$$P_2\left[\mid x_i - x_j \mid < x; L_c\right] = \left(\frac{x}{L_c}\right)\left(\frac{2-x}{2-L_c}\right). \qquad \text{(1-b)}$$

For completely uncorrelated shorts, $L_c = 1$ and $P_2 = 2x - x^2$, and for strongly correlated ones, $L_c \rightarrow 0$ and $P_2 = (x/L_c)$ for $x < L_c$ or $P_2 = 1$, otherwise. The experimental data in Fig. 2(a) can now be replotted for systematic analysis in Fig. 2(b). The CDF for the location of the first short, $x_1^{\{j\}}$, is plotted in Fig. 2 (bottom x-axis, right y-axis). The best fit coefficients for P_1 (Eq. 1a) are ($\alpha = 0.92$, $\beta = 0.04$) compared to the theoretical values of ($\alpha = 1, \beta = 0$). The χ^2- test confirms that the location distribution of the first short is random. Fig. 2(b) also shows (top x-axis, left y-axis) the pair-CDF, P_2, of the distance between the locations of the first and the n-th short ($\mid x_1 - x_n \mid$), with effective correlation distance, $L_c^{(eff)} \geq 0.83$ (see Eq. 1b), confirming that the spatial correlation among the shorts is weak.

The shorts are essentially uncorrelated in location, but are they also uncorrelated in time? That is, does the emergence of one short change the rate of subsequent defect formation, thereby affecting when the next short appears? To check for the *time-correlation*, we compared the n-breakdown Weibull data $W_n = ln\left(-ln(1-F_n)\right)$ (Fig. 3a) with the finite-size percolation model [6] for uncorrelated breakdown

$$-ln(1-F_n) = g_n(\chi) \equiv \chi - ln\left(\sum_{k=0}^{n-1} \chi^k/k!\right) \qquad \text{(2-a)}$$

The scaling parameter $\chi = KA_{ox}t^{\beta}$, where K depends on stress, A_{ox} is the oxide area, and β ($\propto T_{ox}$) is the Weibull slope. For breakdown data for samples having identical thickness, area, and stress, Eq. 2a simplifies to a function of time and number of breakdown events alone, and can be rewritten as

$$\mathbf{g}_n^{-1}[-\ln(1-F_n(t))] = \chi(t) \equiv -ln\left(1 - F_{1\leftarrow n}^{(scaled)}\right). \quad \text{(2-b)}$$

That is, given an experimentally determined F_n ($n > 1$), the series in Eq. 2a can be inverted to obtain $\chi(t)$. The second part of Eq. 2b can then be used to predict the number of samples with at least one short ($F_{1\leftarrow n}^{(scaled)}$) at time t, *provided* the shorts are uncorrelated in time. The coincidence of all the predicted ($1 - F_{1\leftarrow n}^{(scaled)}$) from different n-short distributions, at all times, with the measured ($1 - F_1$), as shown in Fig. 3(b) confirms that, regardless of area and thickness, the shorts are essentially uncorrelated (both spatially and temporally). Note that this result now allows us to measure any one distribution, $F_1(t)$ for example, compute $\chi(t)$, then predict other distributions with Eqs. 2a and 2b - a surprisingly simple functional description for what must be, microscopically, a complicated process of defect generation and short formation.

Redefinition of the reliability specifications

With the breakdown events statistically uncorrelated, consider, in Fig. 4a, the extrapolated Weibull curves for multiple breakdowns (based on Eq. 2a). If an IC contains 100 million transistors (i.e. $N_T = 10^8$), and only 1 in 10^4 IC failures is acceptable, then according to the standard reliability definition, the first transistor failure time, T_1, corresponding to $F_n = 10^{-12}$ ($W_n = -27.6$), would have been the maximum operating lifetime for this technology. T_1 could also have been obtained with area scaling ($A_{IC} = 10^8 A_{ox}$), followed by percentile scaling ($F_n = 10^{-4}, W_n = -9.2$). However, if at least one (statistically-uncorrelated) SBD event per transistor is acceptable[2-7], the TDDB-limited lifetime increases to T_2, at least. In general, if a technology is fault-tolerant up to ($n - 1$) uncorrelated, soft shorts per FET, Eq. 2 indicates that the operating lifetime (t_n) of the IC would increase geometrically over the traditional *fault-intolerant* lifetime (t_1) , i.e.

$$\left(\frac{t_n}{t_1}\right)^{\beta} \approx \frac{C_n}{F_n^{(1-1/n)}}, \quad \text{(2-c)}$$

Note that ***area scaling is no longer as important as it once was***, because SBD in a single transistor no longer defines the failure of the entire IC [Fig. 4b]. Moreover, the importance of percentile scaling is greatly reduced, because the asymptotic slope of the W_n is $n\beta$, not β [Figs. 4a and 4c].

While T_2 (or T_n, if oxides can sustain n breakdown events) now becomes the new definition of TDDB-limited lifetime, other lifetimes, like the *leakage-limited* lifetime, given by

$$T_{leak} = \left[\frac{I_{lmt}}{I_o N_T}\right]^{\frac{1}{\beta}} T_o, \quad \text{(2-d)}$$

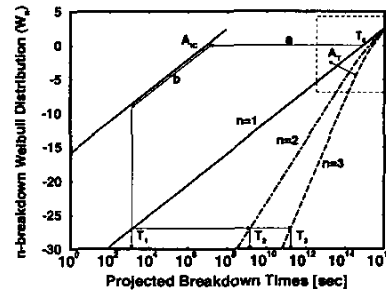

Fig. 4(a): The detrimental effects of area (extrapolated by Eq. 2a) and percentile scaling, both reflecting the reduction of Weibull slope β with oxide thickness, make the projected lifetime T_1 significantly smaller than the median lifetime T_o [dotted window corresponds to the window in Fig. 3(a)]. However, the reduction in operating voltage makes all oxide breakdowns soft, increasing the TDDB-lifetime by orders-of-magnitude, to at least $T_2 \gg T_1$.

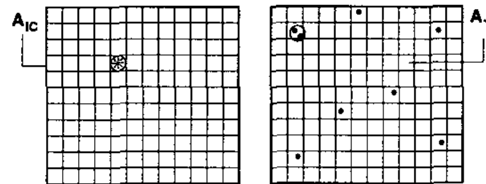

Fig. 4(b): With the standard reliability specification, if a transistor fails, so does the IC (left panel). Such area-scaling makes the TDDB-lifetime, T_1, unacceptably small, especially for large ICs with thin oxides. However, if transistor can sustain at least one SBD each (right panel), and if these SBD events are statically uncorrelated, the size of an IC is no longer important. Rather it is the time-to-multiple SBD within a transistor (T_n), which increases geometrically with the reduction in the transistor area, that controls the TDDB-limited lifetime of a given technology.

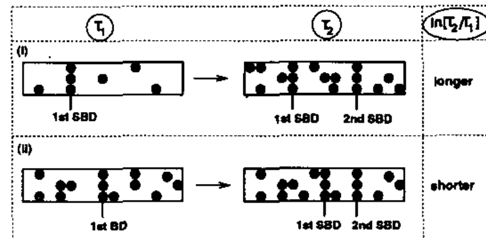

Fig. 4(c): If the 1st-SBD time T_1 is short, as in (i), the sample has not yet accumulated many traps, so that the 2nd SBD at T_2 is delayed (statistically). In contrast, if 1st-SBD at T_1 occurs relatively late, as in (ii), the sample is already full of defects, so that the 2nd SBD occurs relatively quickly. This statistical reduction in $ln(T_2/T_1)$ with larger T_1 causes two-SBD distribution ($W_{n=2}$) to have a higher Weibull slope than one-SBD distribution ($W_{n=1}$), as shown in Fig. 4a. Note that, it is $ln(T_2/T_1)$, not necessarily $| T_2 - T_1 |$, which is reduced, because trap generation itself becomes slower at longer times: $N_{trap} \sim t^{0.4-0.6}$.

could be shorter [10] (I_{lmt}: acceptable level of leakage, I_o: average current per SBD, and T_o: mean failure time.) Similar limits on lifetime could arise from HCI- or NBTI-damage, as well.

Fig. 5(a) shows how the redefinition proposed above modifies existing TDDB reliability limits. In conventional reliability projections, one uses the measured $T_{BD}(V_g)$ to construct curve A, correct for area (curve B), and percentile (curve C) scaling, to determine the 10-year safe operating voltage, V_{max}^{1st-BD}. According to the new definition, however, percentile scaling is weaker (curve D), and area scaling is unnecessary, making the $V_{max}^{(SBD)} \gg V_{max}^{(1st-BD)}$. However, the actual operating voltage should be low enough to guarantee that all breakdowns are soft (Fig. 5b and Ref. [6]), so that

$$V_{max} = min|V_{max}^{(SBD)}, V_{crit}^{(SBD)}| \quad (3)$$

Fig. 6 shows that this criterion dramatically modifies the V_{max} (particularly for PMOS devices) for the technologies operating below 1.2-1.3 volts, making them reliable despite their low Weibull slopes.

Conclusions

In this work, we have established, for the first time, a theoretical framework to study the space- and time- correlation of soft breakdown events. Based on this framework, we have proposed a redefinition of the reliability specifications of ultra-thin, soft broken oxides. This reformulation has allowed us to speculate that oxide reliability may no longer be a concern for IC scaling if soft-breakdown events are statistically uncorrelated, and if at least one soft-breakdown per transistor is acceptable. Our results imply that most modern ICs may have more design margin than previously anticipated, especially for those technologies with operating voltage below 1.3-1.4 volts. Some of this design margins could be traded for higher operating speed (higher voltage) without sacrificing functional reliability. While other difficult challenges to scaling, like power management problems (caused by excessive gate leakage), shallow junction formation, etc. remain and must be addressed effectively [9], our work implies that oxide reliability itself may no longer be considered a fundmental roadblock to continued scaling of silicon ICs.

References:

1. M. Alam *et al.*, ECS Proc. 2000-2, 365.
2. B. Weir *et al.*, ECS Proc. 2002-2, 465.
3. J. Bude *et al.* IEDM Digest, 179, 1998.
4. B. Linder *et al.*, VLSI Digest, 214, 2000,
5. B. Weir *et al.*, IEDM Digest, 437, 1999;
6. M. Alam *et al.*, ITED, **49**, 239, 2002.
7. B. Kaczer *et al.*, IEDM Digest, 553, 2000.
8. R. Degraeve *et al.* IRPS Proc., 360, 2001.
9. P. Packen, Science, **285**, 2079, 1999.
10. K. Okada *et al.*, VLSI Digest, 57, 1999.
11. E. Wu *et al.*, Semi. Sci. Tech., **15**, 425, 2000.

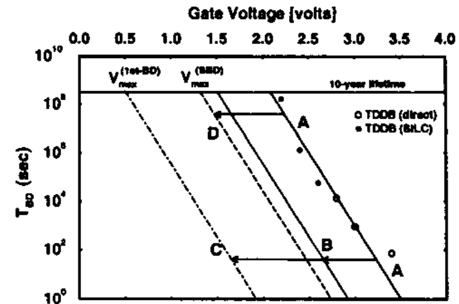

Fig. 5(a): Measured data [2] in curve A are corrected for area (curve B) and percentile (curve C) scaling, to obtain the conventional safe operating voltage, $V_{max}^{(1st-BD)}$. However, adoption of statistically-independent soft-broken oxides makes area-scaling unnecessary and reduces the importance of percentile scaling (curve D). This raises the safe operating voltage significantly.

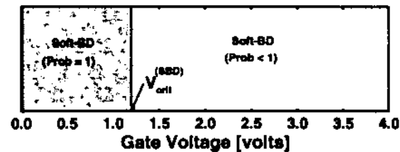

Fig. 5(b): According to the standard theory of SBD phenomena, the ratio of HBD-to-SBD events depends on the power dissipation P through the percolation path [5]. To ensure that all breakdown events are soft, the circuits must operate below $V_{crit}^{(SBD)}$, so that the power dissipation $P_{max}(= V_{crit}^2/R_{min}^{(perc)})$ never exceeds the critical power density P_{crit}. The smaller of the two safe operating voltages from Fig. 5(a) and (b) dictates the maximum operating voltage for a technology. The estimate of $V_{crit}^{(SBD)}$ shown here is based on a conservative value of $P_{max} = 100\ \mu W$.

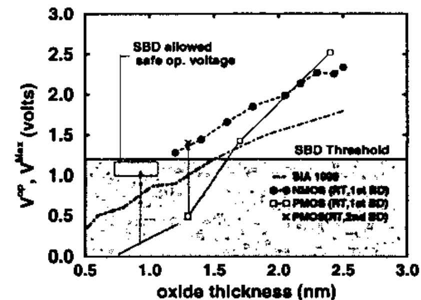

Fig. 6: Plot of the SIA-roadmap along with the safe operating voltage for NMOS and PMOS devices, showing how PMOS reliability becomes a concern for $T_{ox} < 1.8$ nm (open squares), based on reasonable estimates for the pMOS data from the literature[2,11]. However, adoption of statistically-uncorrelated soft-broken oxides raises safe operating voltage significantly, e.g., from 0.5 V to 1.41 V for 1.3 nm oxides (arrow between □ and ×)), although V_{max}=1.1-1.2 V by Eq. 3. Similar improvement in V_{max} for all future technologies is expected, although it may involve multiple soft breakdowns per transistor (multiple-arrows at 0.9 nm here represent transition from $T_1 \rightarrow T_n$ in Fig. 4a.)