Anomalous telegraph noise in small-area silicon metal-oxide-semiconductor field-effect transistors

M. J. Uren, M. J. Kirton, and S. Collins
Royal Signals and Radar Establishment, Great Malvern, Worcestershire WR14 3PS, United Kingdom
(Received 2 November 1987)

In the drain current of submicrometer silicon metal-oxide-semiconductor field-effect transistors, we have observed a new class of random telegraph signal which exhibits anomalous behavior. We discuss the various models that could account for these signals and suggest that they are due to individual Si/SiO_2 interface states which can exist in two or more charge-equivalent, metastable states. We describe one particular signal which is consistent with sequential two-electron capture involving a number of such metastable states at a single defect. We point out that these signals are a source of non-Gaussian noise.

I. INTRODUCTION

In very small-area (<1 μm^2) silicon metal-oxide-semiconductor field-effect transistors (MOSFET's), the fluctuating occupancy of individual Si/SiO_2 interface states generates discrete switching behavior in the drain current.[1,2] The study of the resultant random telegraph signal (RTS) as a function of temperature and gate voltage allows one to build up a detailed picture of the charge-carrier capture and emission processes.[3-5] In this paper, we present a full description and analysis of a new class[6] of RTS which exhibits anomalous, and sometimes very complex, switching behavior (see Figs. 1 and 3). We conclude that the most plausible explanation for this behavior is that it is a direct demonstration of a fundamental property of amorphous materials; namely, the ability of local lattice configurations to exist in two or more metastable states.[7-9] These signals also explain the recent observations of non-Gaussian behavior in the noise properties of small silicon-on-sapphire (SOS) and GaAs resistors.[10,11]

II. TWO-LEVEL ANOMALOUS SIGNALS

Silicon-gate, n-channel MOSFET's with electrical channel dimensions of 0.5×0.75 μm^2 and gate-oxide thickness of 40 nm were employed. Figure 1 depicts the unique switching behavior found in the drain current of one such device. This complex behavior was relatively common; out of a total of 320 RTS's that we observed, 12 were found to be anomalous. Within the time window t_1-t_2 the rapidly switching RTS shows the conventional behavior corresponding to fluctuations in occupancy of an individual interface state.[1,2] Most RTS's have precisely this form for times extending to $\pm\infty$; the times in the up and down states correspond to single-electron capture and emission, respectively.[2,5] Capture times decrease strongly with increasing gate voltage and hence electron concentration, whereas emission times are normally gate-voltage independent or increase with increasing gate voltage. During periods such as t_2-t_3 shown in Fig. 1, the rapid switching completely disappears and the RTS maintains its low level. It thus appears that the fast-switching RTS is modulated in time, with the envelope of modulation itself being an RTS of the same amplitude.

The initial observations of this phenomenon were in two devices operating in weak inversion, indicating that the likely cause was either two traps on the same percolation channel or Coulombic interactions between pairs of traps in close proximity (spatially and energetically). The percolation model is restricted to the weak inversion regime, since it requires inhomogeneous current flow due to partially screened charge centers at the interface. However, the model was quickly discarded since the majority of anomalous signals were subsequently observed in strong inversion where percolation is not a major current-carrying mechanism.

The Coulombic model requires two traps with relatively fast and slow time constants to be colocated and with both their occupancy levels residing close to the Fermi level. The rapid switching shown in Fig. 1 is then due to the fluctuations in occupancy of the fast trap. However, on electron capture into the slow trap the occupancy level of the fast trap will be lifted higher in energy by an amount which can be estimated from simple electrostatics. If the fast trap's occupancy level is moved by several kT the rapid fluctuations will cease and the current will remain at its low level, as in period t_2-t_3, until such time as the slow trap releases its trapped electron and the fast trap is free to fluctuate in occupancy again.

The key feature of the anomalous RTS in Fig. 1 which the Coulombic model has to explain is the presence of two and not three or four current levels. (The only three-level signal that we observed was clearly not the result of Coulombic interaction between pairs of defects and is described later.) During those periods in which the slow trap has captured an electron no observable fluctuations of the fast trap take place, otherwise three current levels would be present. This requires the following set of conditions: either the channel in the immediate vicinity of the fast trap is completely blocked (inversion charge excluded), so that the fluctuations are still occurring but are not observable; or else the occupancy level of the fast trap is moved several kT away from the Fermi level so the trap's occupancy does not fluctuate.

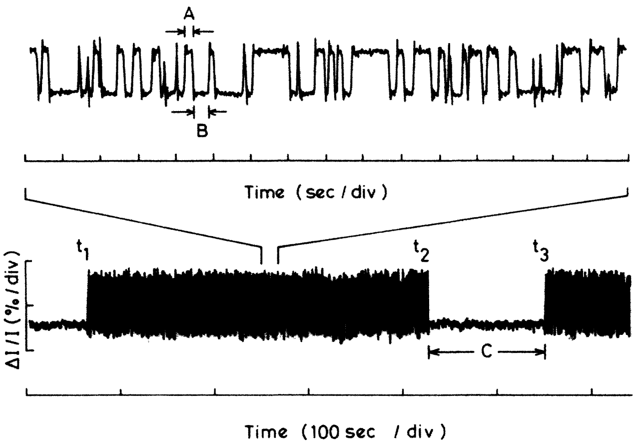

FIG. 1. Fluctuations in current vs time showing a rapidly switching RTS modulated by an envelope of the same amplitude. The upper trace is an expansion of part of the lower trace. $V_G = 1$ V, $V_D = 10$ mV, $I_D = 10.3$ nA, and $T = 293$ K.

Referring first to the channel-blocking model, the important point to note is that on electron capture into a defect there is a localized increase in resistance in the vicinity of the defect; but in practice the channel is never completely blocked. The primary effect is locally to move the inversion-layer electrons away from the interface. In the worst case of weak inversion where there is no screening, evaluating Poisson's equation classically showed that the local free-carrier concentration perpendicular to the interface and directly above the defect fell to a minimum of 21% of its value before trapping. Thus any continuing fluctuation in occupancy would always be visible.

If the effect shown in Fig. 1 were due to one trap moving the occupancy level of a second trap away from the Fermi level, then the distribution of the defects on the interfacial plane is a crucially important parameter. In order to test whether anomalous RTS behavior could arise from the interaction of randomly distributed defects, we performed a simple numerical simulation of the situation corresponding to the device operating in strong inversion. Experimentally it was found that in strong inversion, at a given gate voltage, typically between three and five RTS's were visible in the time window 1 ms–10 s. The simulation was carried out assuming that in a device of dimensions 0.5×0.75 μm^2 five defects were distributed randomly on the interfacial plane. 10 000 simulations were carried out to calculate what percentage of devices would contain defects in sufficiently close proximity to cause some degree of correlation; a separation of 7 nm—or a potential energy shift of 5 meV—was used for this purpose. The simulation assumed a two-dimensional electron gas[12] which is obviously not a complete description at room temperature, but is quite accurate enough for the present purpose (using a classical result gave similar answers). The simulation found about $\frac{1}{40}$ of the number of anomalous signals actually observed. Moreover, the criterion for proximity included those pairs of defects which would show intermediate behavior: i.e., some RTS's would show three levels due to a slight modification of one trap's time constant caused by occupation of a rather distant second trap. So we can conclude that if the anomalous behavior of Fig. 1 were the result of pairs of defects, then it would require that the pairs be grouped at specific sites and not randomly distributed. In order to explain the complete absence of the third level, the screened Coulomb potential must shift the energy level of the other defect by several kT. We estimate that the defect separation must be less than ~ 2 nm to achieve this.

Clustering of defects in the oxide at some inactive defect would appear to be a logical cause of the high number of anomalous signals seen. Since there are stringent requirements on the maximum allowable defect separation (~ 2 nm), this necessitates that the defects decorate inactive point defects; clustering at line and planar defects is ruled out because such clustering does not restrict the separation of defects to less than the critical value. Clustering of heavy-metal ions has been observed at the Si/SiO_2 interface in intentionally contaminated samples,[13] with the clustering occurring at a small number of nucleation sites; but in good device-quality interfaces, conductance measurements do not show anything other than a random distribution of charge (although this technique would not normally be sensitive to clustering on length scales much less than the oxide thickness). Recent experiments using the scanning tunneling microscope have shown defect clustering on submonolayer oxide surfaces,[14] but this is far removed from our situation.

The requirement of very closely spaced pairs of defects prompts the question: Should the composite defect be considered a single defect, arising out of the union of two defect potentials, and leading to a single occupancy level? Whilst it is not possible to answer this question in detail—since current theoretical and experimental knowledge of defects in SiO_2 and its interfaces is limited—it is clear that an explanation based on two independent traps is unlikely to be complete. The two-trap model has to be very carefully molded: it is only consistent with the data if the defects are clustered around a point defect and their separation does not exceed ~ 2 nm, but they must be far enough apart to prevent hybridization.

A simpler explanation consistent with our observations is that the signals result from a single defect with two reconstruction modes (metastable states) available for the filled trap. This hypothesis immediately accounts for the fact that the amplitudes of the underlying RTS and its envelope are equal and the total absence of a third level. Moreover, it is particularly appropriate in view of the established evidence of metastability in glassy systems[7–9] and recent observations of significant electron-lattice coupling at individual Si/SiO_2 defect states.[1,5]

Figure 2 shows schematic configuration-coordinate diagrams of two models which exhibit metastability; the models differ only in the way the various states intercommunicate. Rapid electron capture and emission proceeds via total-energy minimum α_1 and accounts for section $t_1 - t_2$ of Fig. 1. In the model shown in Fig. 2(a), after

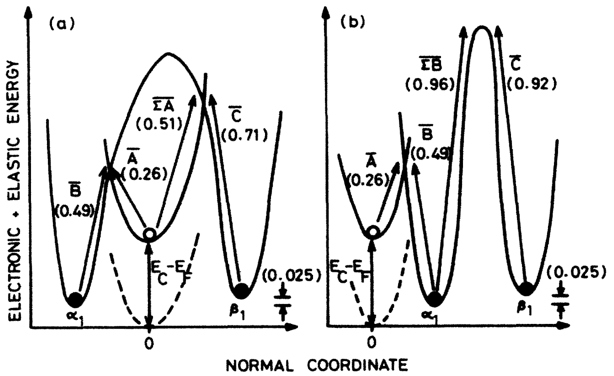

FIG. 2. Two possible configuration-coordinate diagrams for the defect whose RTS is shown in Fig. 1. The dashed curve shows the empty trap before the creation of a free electron in the conduction band: ○ labels the empty trap plus a free electron; ● marks the metastable states containing one trapped electron. The observed transitions are labeled with the corresponding average times from Fig. 1. (a) Both metastable states are in communication with the inversion layer. The activation energies in eV are given in parentheses (Ref. 5). These are also the activation energies one would obtain using a Coulombic interaction interpretation of the data. (b) Only α_1 is able to capture an electron directly. Activation energies for the transformations $\alpha_1 \leftrightarrow \beta_1$ were evaluated assuming $\tau = \tau_0 \exp(E/kT)$.

electron capture into β_1 it will take a time, $\overline{C}$ ($\overline{C} \gg \overline{A}, \overline{B}$), for the filled trap to reemit the electron. During this time interval no switching of the RTS will occur, and the drain current will remain fixed at its low level as in period $t_2 - t_3$. For the model of Fig. 2(b), the transformation from α_1 into state β_1 should be thermally activated and roughly gate-voltage independent, with $\overline{C}$ representing the time taken for the physical reconfiguration of the one electron state from β_1 to α_1. Referring to Fig. 1, we see that the time for electron capture into α_1 is given by the average of time A, $\overline{A}$; the emission time from α_1 is $\overline{B}$. For the model in Fig. 2(a), the capture time into β_1, $\overline{\sum A}$, can be obtained by summing the time A over an average time window $t_1 - t_2$; alternatively, for the model in Fig. 2(b) the transformation time from α_1 to β_1 is given by $\overline{\sum B}$.

Using the gate-voltage dependence[1] of $\overline{A}$ and $\overline{\sum A}$, we found that the metastable states α_1 and β_1 were located the same distance into the oxide. Using the grand partition function[15] to evaluate equilibrium occupancies, we found that β_1 lay above α_1 by 0.025 eV. In fact, both metastable models shown in Fig. 2 were consistent with our data and we were unable to distinguish between them. The important point to note, however, is that the concept of metastability provides a simple yet elegant explanation for the anomalous RTS of Fig. 1, independent of the transformation mechanism.

III. THREE-LEVEL ANOMALOUS BEHAVIOR

Up to this point, we have discussed the behavior characteristic of the majority of anomalous RTS's that we observed. We would now like to consider one particular RTS which showed complex and interesting behavior (see Fig. 3). The first point to note is that it is a three-level signal with the separation between levels 0 and 1 equal to the separation between levels 1 and 2. In addition, the rapid switching represented by times U and V occurs only between levels 1 and 2 showing that the RTS is not just a straightforward superposition of two distinct and independent RTS's. These facts taken together suggest that the signal represents a sequential two-electron capture process: transition $0 \to 1$ being capture of the first electron; transition $1 \to 2$ being capture of the second electron. Thus it is a candidate for the Coulombic effect alluded to earlier.

To address this possibility, we investigated the gate-voltage dependence of the various time constants (see Fig. 4). $\overline{P}$ and $\overline{U}$ are both strongly gate-voltage dependent with the same slope. This demonstrates that they represent straightforward single-electron capture: $\overline{P}$ of the first electron; $\overline{U}$ of the second electron. If this signal were the result of Coulombic effects, then $\overline{P}$ would represent capture into the relatively fast state with the slow state empty, and $\overline{U}$ capture into the relatively fast state with the slow state full. Since the effect of filling the slow trap would be to lengthen the time for capture into the fast trap (by moving the carriers away from the Si/SiO_2 interface hence reducing the local carrier concentration and cross section), it is required that $\overline{U} > \overline{P}$. Since in fact $\overline{U} < \overline{P}$ for all gate voltages, it was not possible to explain the signal by simple Coulombic interaction between two traps.

We shall now show that the behavior can be explained by a model involving two-electron capture at a single defect exhibiting metastability. Immediately after capture of the first electron and before a transition to the charge-2 level takes place, the RTS spends a period of time $\overline{S}$ in the charge-1 level. At high gate voltages, $\overline{S}$, has the same gate-voltage dependence as both $\overline{P}$ and $\overline{U}$, whereas at low gate voltages it becomes independent of gate voltage. These two facts are consistent with $\overline{S}$ representing the capture of a second electron in competition with emission of the first electron.

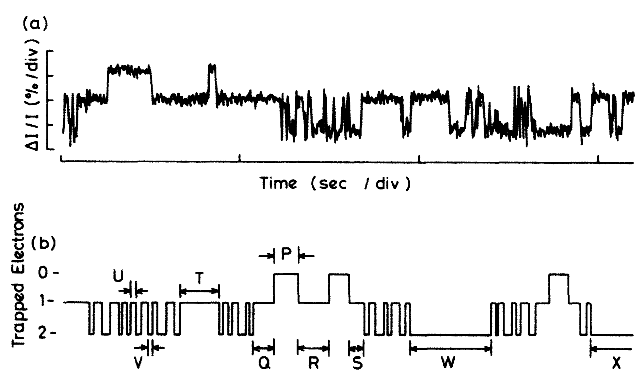

FIG. 3. (a) Complex RTS generated by a single defect which shows one- and two-electron capture. $T = 290$ K, $V_G = 0.6$ V, $V_D = 20$ mV, and $I_D = 195$ pA. (b) Schematic RTS showing all the features which were observed for this defect.

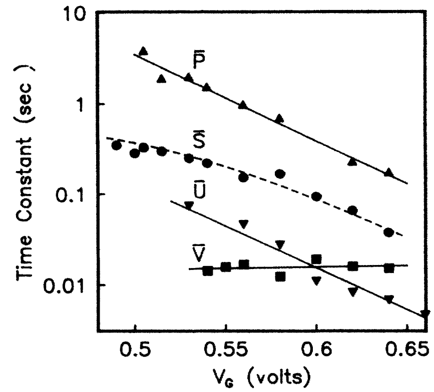

FIG. 4. Gate-voltage dependence of representative times $\bar{P}$, $\bar{S}$, $\bar{U}$, and $\bar{V}$ from the RTS shown in Fig. 3. The solid lines are regression lines and the dashed line is a fit assuming simple competition between emission and capture in the θ_1 state.

Turning now to the other time constants observed in the RTS of Fig. 3, we found that $\bar{V}$ was independent of gate voltage and thus corresponded to a simple single-electron emission process. We were not able to study W and X in any detail: W was quite a rare event lasting around 0.5 sec; and X turned off all fluctuations for several hours, making all measurements of this complex defect rather difficult.

Based on this information, an appropriate configuration-coordinate diagram for this defect is shown in Fig. 5. The energy zero has been chosen as the defect in the charge-0 level with two electrons at the Fermi level. Taking one of these electrons from the Fermi reservoir and placing it in the conduction band increases the total energy of the system by an amount equal to $E_C - E_F$. Metastable minimum θ_1 (one electron captured), for example, is separated from the energy zero by an energy $(E_C - E_F) - [E_C - E_{\theta_1}(1/0)] = E_{\theta_1}(1/0) - E_F$; minimum θ_2 (two electrons captured) is separated by an energy $[E_{\theta_1}(1/0) - E_F] + [E_{\theta_2}(2/1) - E_F]$.

The charge-1 level consists of two metastable states, θ_1 and λ_1. Capture of the first electron from the inversion layer, with time constant $\bar{P}$, takes place into the θ_1 state and not into the λ_1 state. Capture of the second electron initially takes place from θ_1 into θ_2 with time constant $\bar{S}$. Thereafter, fluctuations in occupancy of θ_2 take place: via $\theta_2 \leftrightarrow \lambda_1$ with relatively fast time constants $\bar{U}$ and $\bar{V}$; and via $\theta_2 \leftrightarrow \theta_1$ with slower times $\bar{T}$ $(=\bar{S})$ and $\overline{\sum V}$. An important observation is the fact that direct, thermally activated (and gate-voltage-independent) transformation between the metastable states θ_1 and λ_1 of the charge-1 level was not observed. Instead, transformation between the two always took place via an intermediate state θ_2 of different charge.

The energies of the various occupancy levels were evaluated using the formalism of the grand partition function,[15] assuming all degeneracies were equal. The energy levels are shown in the inset to Fig. 5. The important point to note is that the occupancy levels $E_{\theta_1}(1/0)$ and

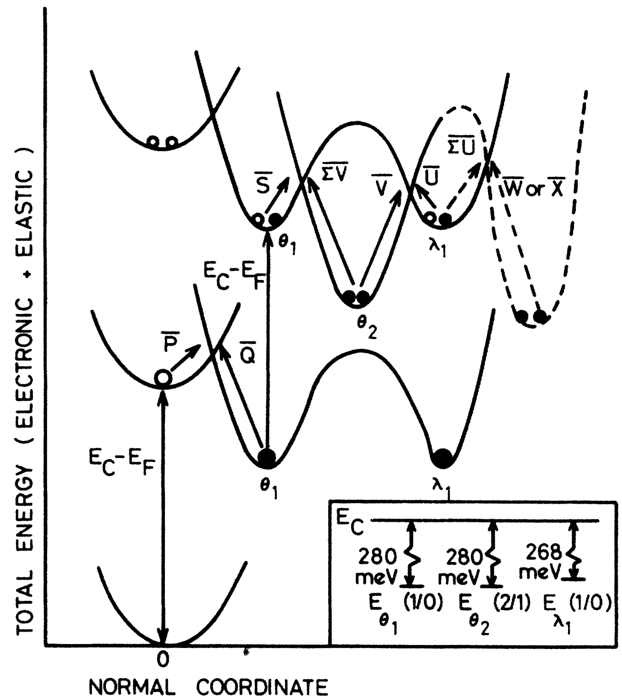

FIG. 5. Schematic configuration-coordinate diagram for the defect whose RTS is shown in Fig. 3. The diagram shows the total energy of the defect for the three cases of 0, 1, and 2 electrons removed from the reservoir (at energy E_F): ○ indicates a free electron in the conduction band and ● a trapped electron. The arrows show the various transitions we have identified; the dashed section is speculative and incomplete. Varying the gate-voltage changes $E_C - E_F$, the relative positions of the three sets of curves, and hence the occupancy of the defect. The inset gives the trap energy levels we evaluated.

$E_{\theta_2}(2/1)$ were found to be degenerate to within a few meV. This means that as the Fermi level crosses the occupancy level $E_{\theta_1}(1/0)$—and the charge state changes from 0 to 1—it also crosses the occupancy level $E_{\theta_2}(2/1)$ and so the defect is immediately capable of capturing a second electron. This very nearly corresponds to a negative-U system in which the occupancy level $E(2/1)$ would lie below $E(1/0)$. In such a system, the occupancy of the defect would change directly from 0 to 2 missing the charge-1 state. The extra electron-lattice interaction accompanying the capture of the second electron offsets the additional Coulombic interaction energy.[16]

We have hitherto referred to the charge levels of the defect as 0, 1, and 2, and inspection of Fig. 3 shows that the amplitudes of the $0 \rightarrow 1$ and $1 \rightarrow 2$ transitions are nearly equal. This can be explained by a defect whose charge state changes from + to 0 and then 0 to −; the occupancy levels would then be $E(0/+)$ and $E(-/0)$. If we denote the average local free-carrier concentrations surrounding the defect as n^+, n^0, and n^-, then, due to the exponential dependence of carrier concentration on surface potential, $n^+/n^0 = n^0/n^-$. The local conductances obey the same equality. Using a simple resistive network model or effective-medium theory, one can then show that (for small changes) $\Delta R/R$ for the $+\rightarrow 0$ transition is equal and opposite to the change for $0\rightarrow -$. The charge assignments and negative-U properties would then be consistent with the defect being, for example, some form of dangling hybrid.

IV. NON-GAUSSIAN NOISE

We would like to consider the incidence of metastability and its implications for noise statistics. Our observed figure of $\sim 4\%$ of defects exhibiting anomalous behavior is very much a lower bound. The true fraction is probably considerably greater. Our experimental setup only measures RTS's whose mark-space ratio is close to unity, and thus we are only capable of detecting metastable states if all the energy levels lie within a few kT of the Fermi level.[2] In practice, one expects a wide range of energy-level separations and barrier heights.[7]

Restle *et al.*[10,11] have shown that in small SOS and GaAs resistors the integrated noise power in a given octave exhibits low-frequency amplitude modulations as a function of time. In GaAs, the modulations were themselves found to exhibit a $1/f$ spectrum. The RTS shown in Fig. 1 possesses all the required characteristics to be the origin of this non-Gaussian behavior. The modulating envelope accounts for the low-frequency modulations of the integrated power as we have verified in recent numerical simulations.[17] Moreover, the observed $1/f$ spectrum in the amplitude variations must arise from a wide distribution of time constants of the modulating envelopes. We saw a spread of envelope time constants ranging from milliseconds to days.

V. CONCLUSION

We have discovered anomalous random telegraph signals in the drain current of small-area silicon MOSFET's. These defects have hitherto not been visible using conventional techniques which only measure average properties. The signals can be explained by Coulombic interactions between clustered defects; but a more realistic and simple explanation is that they are due to charge trapping into individual metastable SiO_2 interface states. The wealth of information available from these small devices has allowed us to identify in one particularly complex signal, two-electron capture and negative-U-like behavior at a single defect.

ACKNOWLEDGMENT

We would like to thank R. Lambert of the GEC Hirst Research Center (Wembley, United Kingdom) for the devices.

[1]K. S. Ralls *et al.*, Phys. Rev. Lett. **52**, 228 (1984).
[2]M. J. Uren, D. J. Day, and M. J. Kirton, Appl. Phys. Lett. **47**, 1195 (1985).
[3]L. D. Jackel *et al.*, in *Proceedings of the 17th International Conference on the Physics of Semiconductors, San Francisco, 1984,* edited by J. D. Chadi and W. A. Harrison (Springer, New York, 1985), p. 221.
[4]C. T. Rogers and R. A. Buhrman, Phys. Rev. Lett. **55**, 859 (1985).
[5]M. J. Kirton and M. J. Uren, Appl. Phys. Lett. **48**, 1270 (1986).
[6]M. J. Uren, Bull. Am. Phys. Soc. **32**, 394 (1987).
[7]P. W. Anderson, B. I. Halperin, and C. M. Varma, Philos. Mag. **25**, 1 (1972).
[8]W. A. Phillips, J. Low Temp. Phys. **7**, 351 (1972).
[9]L. Guttman and S. M. Rahman, Phys. Rev. B **33**, 1506 (1986).
[10]P. J. Restle, R. J. Hamilton, M. B. Weissman, and M. S. Love, Phys. Rev. B **31**, 2254 (1985).
[11]P. J. Restle *et al.*, Phys. Rev. B **34**, 4419 (1986).
[12]F. Stern and W. E. Howard, Phys. Rev. **163**, 816 (1967).
[13]E. H. Nicollian and J. R. Brews, *MOS Physics and Technology* (Wiley, New York, 1982).
[14]R. H. Koch, Bull. Am. Phys. Soc. **32**, 394 (1987).
[15]C. Kittel and H. Kroemer, *Thermal Physics* (Freeman, San Francisco, 1980).
[16]P. W. Anderson, Phys. Rev. Lett. **34**, 953 (1975).
[17]M. J. Kirton, M. J. Uren, and S. Collins, Appl. Surf. Sci. **30**, 148 (1987).