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, as we noted in section 3.2.3. 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 falls to a minimum of 22% of its value before trapping. Thus any continuing fluctuation in occupancy would always be visible, albeit of reduced amplitude.

If the effect shown in figure 38 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. To test whether complex RTS behaviour 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 RTSs were visible in the time window from 1 ms to 100 s. The simulation was carried out assuming that in a device of dimensions $0{\cdot}5\,\mu m \times 0{\cdot}75\,\mu m$ 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, corresponding to a potential-energy shift of 5 meV, was used for this purpose. The simulation assumed a two-dimensional electron gas (Stern and Howard 1967), 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 1/40th of the number of complex signals actually observed. Moreover, the criterion for proximity included those pairs of defects that would show intermediate behaviour: that is, some RTSs 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 behaviour of figure 38 were the result of pairs of defects then it would require that the pairs be grouped at specific sites and not randomly distributed. 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 about 2 nm to achieve this.

Clustering of defects in the oxide at some inactive defect would appear to be a logical cause of the number of complex signals seen. Since there are stringent requirements on the maximum allowable defect separation (about 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 (Nicollian and Brews 1982), 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 tunelling microscope have shown defect clustering on submonolayer oxide surfaces (Koch 1987), but this is far removed from the situation investigated by us.

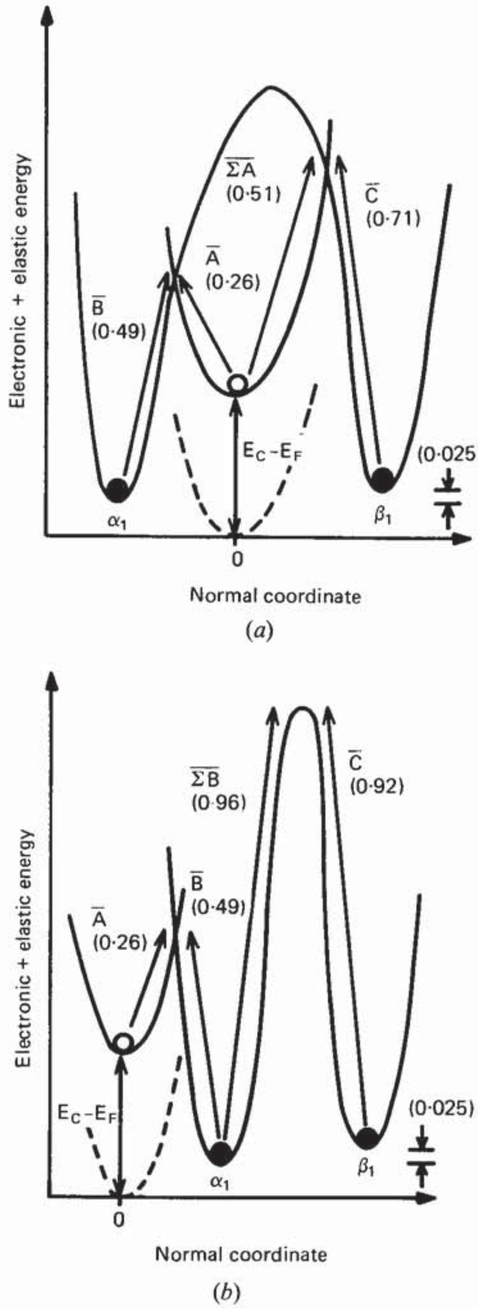
The requirement of very closely spaced pairs of defects prompts the question as to whether the composite defect should be considered a single defect, arising out of the union of two defect potentials, and leading to a single occupancy level? While it is not possible to answer this question in detail, since the present 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 moulded: it is only consistent with the data if the defects are clustered around a point defect and their separation does not exceed about 2 nm, but they must be far enough apart to prevent significant wavefunction hybridization.

An alternative explanation consistent with the 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 (Anderson *et al.* 1972, Phillips 1972, Guttman and Rahman 1986) and recent observations of significant electron–lattice coupling at individual $Si:SiO_2$ defect states (section 5). This is the mechanism advocated by Rogers and Buhrman (1985) to explain their two-level RTS measurements (section 5.2).

Figures 39 (*a*) and (*b*) show schematic configuration-coordinate diagrams of two models that exhibit metastability; the models differ only in the way the various states intercommunicate. The total-energy zero in both figures corresponds to the empty trap with a free electron residing at the Fermi level. In the model shown in figure 39 (*a*), model (*a*), the two metastable states of the filled trap are denoted by α_1 and β_1, and capture of an electron from the inversion layer takes place directly into either α_1 or β_1. In the model of figure 39 (*b*), model (*b*), capture from the inversion layer takes place only into α_1; thereafter, the defect periodically flips over into total-energy minimum β_1. (These diagrams are in a sense analogues of the two Coulombic models described earlier. With strong electron–lattice interactions involved, the differences between the metastable and Coulombic models are rather small.)

For both models (*a*) and (*b*) rapid electron capture and emission proceeds via total-energy minimum α_1 and accounts for the section from t_1 and t_2 of figure 38. After electron capture into β_1 it will take a time $\overline{C}(\overline{C} \gg \overline{A}, \overline{B})$ for the filled trap to re-emit the electron; during this time interval, no switching of the RTS will occur. In addition, total-energy minimum β_1 traps the electron at the same physical location in the oxide as does α_1. Thus the drain current will remain fixed at the low level of the period from t_1 to t_2; this accounts for the section from t_2 to t_3 of figure 38. (This is not necessarily always true: see section 6.6.)

Models (*a*) and (*b*) were both found to be consistent with the gate-voltage dependent data presented in tables 7 and 8. For model (*a*), the capture time $\overline{A}$ from the inversion layer into minimum α_1 is strongly gate-voltage dependent; the emission time from α_1, $\overline{B}$, is only weakly dependent on gate voltage; the capture time from the inversion layer into β_1, $\overline{\Sigma A}$, is strongly dependent on gate voltage; the emission time from $\beta_1 \overline{C}$, is weakly dependent on gate voltage; and the capture times into α_1 and β_1 have the same gate-voltage dependences, locating α_1 and β_1 the same distance into the oxide as required. Similarly for model (*b*), the transformation times from α_1 to β_1, $\overline{\Sigma B}$, and its reverse process β_1 to α_1, $\overline{C}$, are both gate-voltage-independent as required. It is thus not possible to discard either model, since for this RTS $\overline{B}$ shows little dependence on



(*a*)

(*b*)

Figure 39. Two possible models for the defect whose RTS is shown in figure 38. 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 labelled with the corresponding average times from figure 38. (*a*) Both metastable states are in communication with the inversion layer. The activation energies in eV are given in parentheses. 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 $\tau = \tau_0 \exp(E/kT)$. After Uren *et al.* (1988).

Table 7. The dependence on gate voltage of the times $\bar{A}$ and $\bar{B}$ for the RTS shown in figure 38. f has been calculated from the relationship $f = \bar{B}/(\bar{A} + \bar{B})$. $T = 321{\cdot}8$ K and $V_D = 10$ mV.

V_G (V)	I_D (nA)	$\bar{A}$ (s)	$\bar{B}$ (s)	f
1·02	18·87	0·0186	0·0214	0·535
1·03	19·89	0·0156	0·0207	0·570
1·06	25·96	0·0098	0·0208	0·679
1·08	29·85	0·0074	0·0196	0·725

Table 8. Gate-voltage dependence of the modulating envelope of the RTS shown in figure 38. $\overline{t_{12}}$ is the mean time t_1 to t_2 and $\overline{t_{23}}$ the mean time t_2 to t_3, $\bar{C}$. The mean time spent in the up state, $\overline{\Sigma A}$, during time interval $\overline{t_{12}}$ has been calculated from the relation $\overline{\Sigma A} = (1 - f)\overline{t_{12}}$. The mean time spent in the down state, $\overline{\Sigma B}$, during time interval $\overline{t_{12}}$ has been calculated from the relation $\overline{\Sigma B} = f\overline{t_{12}}$. $T = 321{\cdot}8$ K and $V_D = 10$ mV.

V_G (V)	$\overline{t_{12}}$ (s)	$\overline{t_{23}}$ (s)	$\overline{\Sigma A}$ (s)	$\overline{\Sigma B}$ (s)
1·02	47·74	10·49	22·19	25·54
1·03	42·85	10·43	18·43	24·42
1·06	41·46	12·18	13·31	28·15
1·08	31·76	11·16	8·74	23·03

gate voltage. The important point to note is that the concept of metastability provides a simple yet elegant explanation for the complex RTS of figure 38, independent of transformation mechanism.

We will now use the grand partition function (section 2.5) to investigate the relative separation of total-energy minima α_1 and β_1 and their gate-voltage dependences. Our nomenclature will be as follows: $E_{\alpha_1}(1/0)$ and $E_{\beta_1}(1/0)$ will denote the occupancy levels of the minima α_1 and β_1; these occupancy levels will be assumed to have the same degeneracy $\gamma(1)$; the energy zero of the system will correspond to the defect in state 0 with the free electron at the Fermi reservoir. Then for either of the models shown in figures 39 (*a*) and (*b*)

$$Z_G = \gamma(0) + \gamma(1)\exp\left[-\frac{E_{\alpha_1}(1/0) - E_F}{kT}\right] + \gamma(1)\exp\left[-\frac{E_{\beta_1}(1/0) - E_F}{kT}\right]. \tag{6.1}$$

The probability of finding the defect in state 0 is

$$p(0) = \gamma(0)/Z_G, \tag{6.2}$$

and the probability of finding the defect in the α_1 minimum is

$$p_{\alpha_1}(1) = \gamma(1)\exp\left[-\frac{E_{\alpha_1}(1/0) - E_F}{kT}\right]\bigg/Z_G. \tag{6.3}$$

The probability $p_{\beta_1}(1)$ is as above with β_1 replacing α_1. Thus

$$\frac{p_{\alpha_1}(1)}{p_{\beta_1}(1)} = \exp\left[\frac{E_{\beta_1}(1/0) - E_{\alpha_1}(1/0)}{kT}\right]. \tag{6.4}$$

On the basis of the data presented in tables 7 and 8, if we take $V_G = 1{\cdot}06\,\mathrm{V}$ as a representative data point then

$$p_{\alpha_1}(1) = \frac{\overline{\Sigma B}}{\overline{\Sigma A} + \overline{\Sigma B} + \overline{C}} = \frac{\overline{\Sigma B}}{\overline{t_{12}} + \overline{C}} = 0{\cdot}525,$$

$$p_{\beta_1}(1) = \frac{\overline{C}}{\overline{t_{12}} + \overline{C}} = 0{\cdot}227.$$

Then, using equation (6.4), we find

$$\exp\left\{\frac{[E_{\beta_1}(1/0) - E_{\alpha_1}(1/0)]}{kT}\right\} = \frac{0{\cdot}525}{0{\cdot}227}$$

and thus $E_{\beta_1}(1/0) - E_{\alpha_1}(1/0) \approx 25\,\mathrm{meV}$; this separation of the two minima is shown in figures 39 (*a*) and (*b*).

The formalism of sections 5.1.1 and 5.1.2 was used to determine the activation energies for model (*a*); the values obtained are shown in parentheses in figure 39 (*a*). For model (*b*) the activation energies of the transformations $\alpha_1 \leftrightarrow \beta_1$ were calculated assuming $\tau = \tau_0 \exp(E/kT)$. These activation energies are exactly what one would calculate for the two Coulombic models.

To end this section, we should like to consider the incidence of metastability and its implications for noise statistics. Our observed figure of about 4% of defects exhibiting metastability must be considered to be very much a lower bound. Experimentally, one only measures RTSs whose mark-space ratio is close to unity. Thus we were only capable of detecting metastable states (α_1, β_1, etc.) if they resided within a few kT of the Fermi level. In practice, one would expect a wide range of barrier heights separating the metastable minima and a wide distribution of minimum-energy separations (Anderson *et al.* 1972, Phillips 1972). Further evidence for this comes from the fact that a number of RTSs were visible for only a few minutes or so before disappearing; they either did not reappear or did so only after a few days.

As we discussed in section 6.1, Restle *et al.* (1985, 1986) 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 figure 38 possesses all the required characteristics to be the origin of this non-Gaussian behaviour. The modulating envelope accounts for the low-frequency modulations of the integrated power, as has been verified in numerical simulations (Kirton *et al.* 1987). 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.

6.3. *Complex telegraph noise in MOSFETs* (*II*): *Three-level signals*

Up to this point, we have discussed the behaviour characteristic of the majority of complex RTSs that we observed. We should now like to consider one RTS that showed particularly interesting behaviour: see figure 40. 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 RTSs. These facts taken together suggest that the signal represents a sequential two-electron

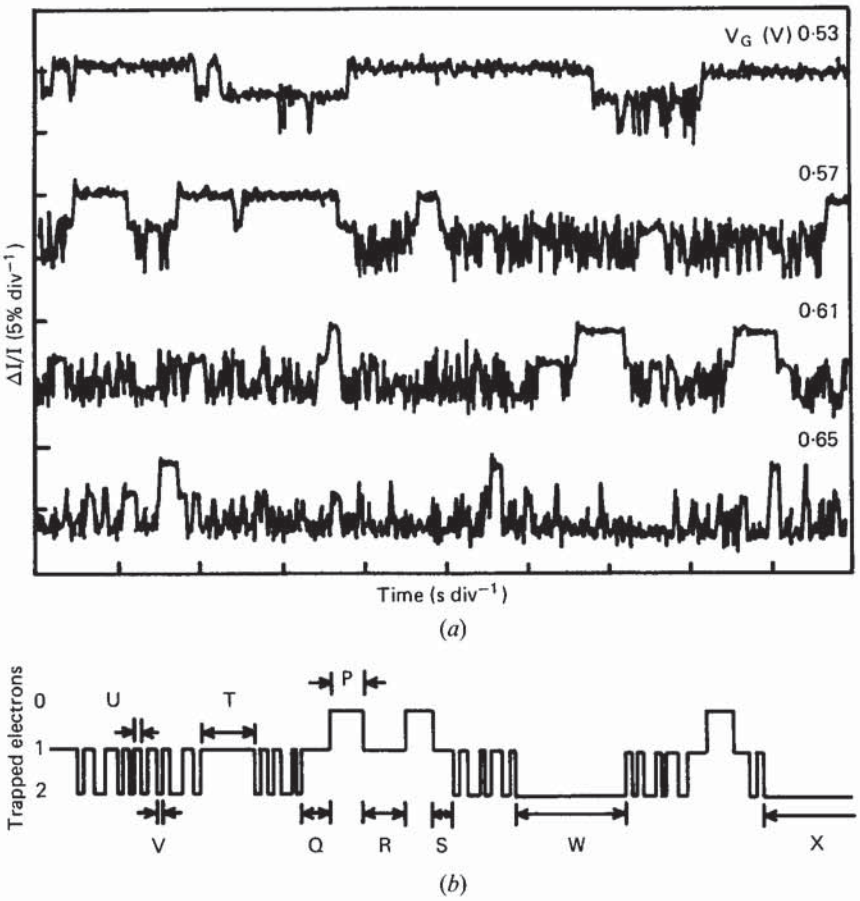


Figure 40. Gate-voltage dependence of a complex RTS generated by a single defect, which shows one- and two-electron capture. $T = 293$ K, $V_D = 20$ mV. $V_G, I_D = 0{\cdot}53$ V, 47 pA; 0·57 V, 97 pA; 0·61 V, 229 pA; 0·65 V, 519 pA. (*b*) Schematic RTS showing all the features that were observed for this defect. After Kirton *et al.* (1988).

capture process: transition 0 → 1 being capture of the first electron and transition 1 → 2 being capture of the second electron. Thus it is a candidate for the Coulombic effect where there is a modification of one defect energy level by the occupation of another.

To address this possibility, we investigated the gate-voltage dependence of the various time constants: see figure 41. $\bar{P}$ and $\bar{U}$ are both strongly gate-voltage-dependent, with the same slope. This demonstrates that they represent straightforward single-electron capture: $\bar{P}$ of the first electron; $\bar{U}$ of the second electron. If this signal were the result of Coulombic effects then $\bar{P}$ would represent capture into the relatively fast state with the slow state empty, and $\bar{U}$ 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 and hence reducing the local carrier concentration and cross-section), it is required that $\bar{U} > \bar{P}$. Since in fact $\bar{U} < \bar{P}$ for all gate voltages, we could not explain the signal by simple Coulombic interaction between two traps.

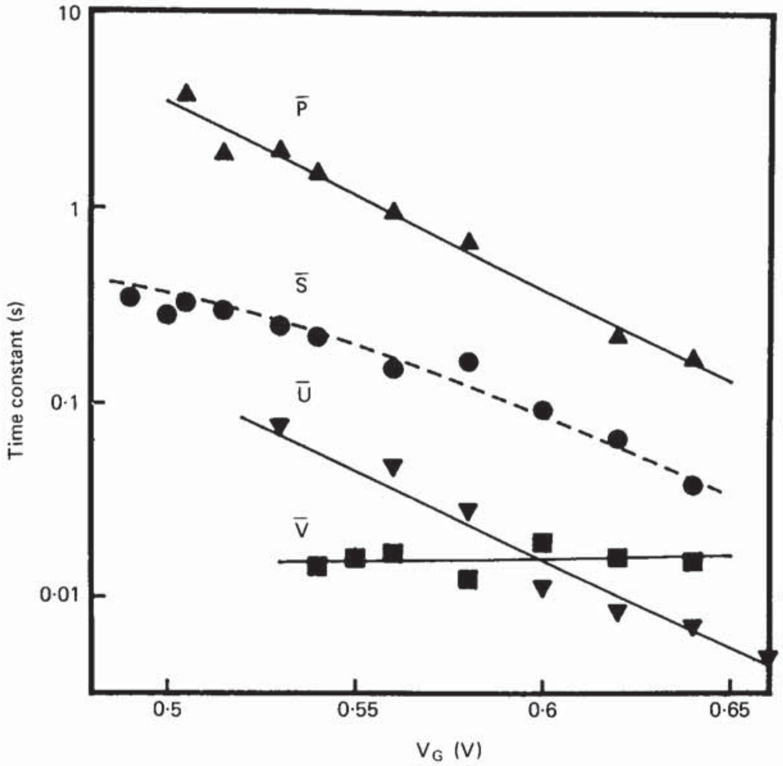


Figure 41. Gate-voltage dependence of representative times $\overline{P}$, $\overline{S}$, $\overline{U}$ and $\overline{V}$ for the RTS shown in figure 40. The solid lines are regression lines and the dashed line is a fit assuming simple competition between emission and capture in the θ_1 state. After Uren *et al.* (1988).

However, the behaviour 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 $\bar{S}$ in the charge-1 level. At high gate voltages, $\bar{S}$ has the same gate-voltage dependence as both $\bar{P}$ and $\bar{U}$, whereas at low gate voltages it becomes independent of gate voltage. These two facts are consistent with $\bar{S}$ representing the capture of a second electron in competition with emission of the first electron, with the shorter of the two times dominating.

Turning now to the other time constants observed in the RTS of figure 40, 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 s; and after a few minutes *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 figure 42. 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 θ1 (one

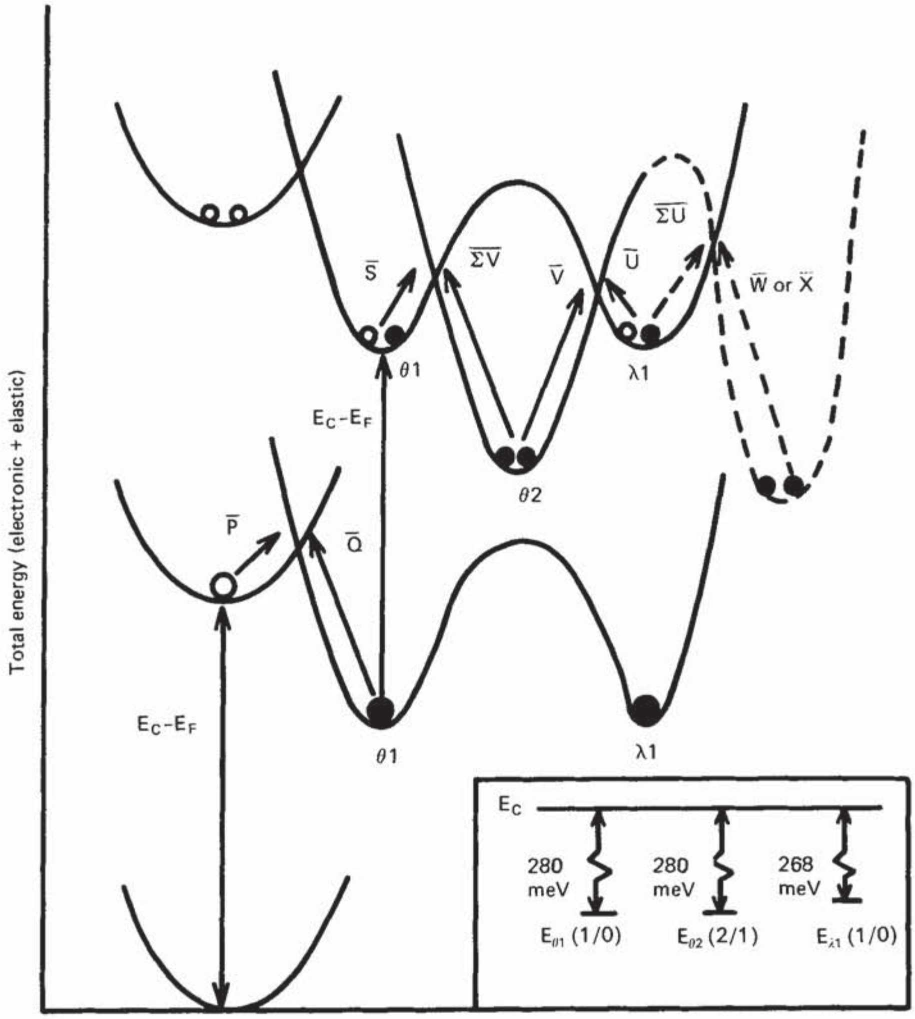


Figure 42. Schematic configuration-coordinate diagram for the defect whose RTS is shown in figure 40. The diagram shows the total energy of the defect for the three cases of 0, 1 and 2 electrons removed from the reservoir (at energy E_F): ○ indicates a free electron in the conduction band and ● a trapped electron. The arrows show the various transitions 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. From Uren *et al.* (1988).

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 θ2 (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, θ1 and λ1. Capture of the first electron from the inversion layer, with time constant $\bar{P}$, takes place into the θ1 state and not into the λ1 state. Capture of the second electron initially takes place from θ1 into θ2, with time constant $\bar{S}$. Thereafter fluctuations in occupancy of θ2 take place: via θ2 ↔ λ1 with relatively fast time constants $\overline{U}$ and $\overline{V}$; and via θ2 ↔ θ1 with slower times $\overline{T}(=\overline{S})$ and $\overline{\Sigma V}$. An important observation is the fact that direct thermally activated (and hence gate-voltage-independent) transformation between the metastable states θ1 and λ1 of the charge-1 level was not observed. Instead, transformation between the two always took place via an intermediate state θ2 of different charge.

The energies of the various occupancy levels were evaluated using the formalism of the grand partition function of section 2.5 along the lines previously described for the metastable defects of section 6.2. Confining our attention to the states θ1, λ1 and θ2 and assuming that all degeneracies are equal to γ, we can write

$$\begin{aligned} Z_{\mathrm{G}} &= \gamma + \gamma \exp\left[-\frac{E_{\theta 1}(1/0) - E_{\mathrm{F}}}{kT}\right] + \gamma \exp\left[-\frac{E_{\lambda 1}(1/0) - E_{\mathrm{F}}}{kT}\right] \\ &\quad + \gamma \exp\left[-\frac{E_{\theta 1}(1/0) + E_{\theta 2}(2/1) - 2E_{\mathrm{F}}}{kT}\right] + \dots, \\ p(0) &= \gamma/Z_{\mathrm{G}}, \\ p_{\theta 1}(1) &= \gamma \exp\left[-\frac{E_{\theta 1}(1/0) - E_{\mathrm{F}}}{kT}\right]\Big/ Z_{\mathrm{G}}, \\ p_{\lambda 1}(1) &= \gamma \exp\left[-\frac{E_{\lambda 1}(1/0) - E_{\mathrm{F}}}{kT}\right]\Big/ Z_{\mathrm{G}}, \\ p_{\theta 2}(2) &= \gamma \exp\left[-\frac{E_{\theta 1}(1/0) + E_{\theta 2}(2/1) - 2E_{\mathrm{F}}}{kT}\right]\Big/ Z_{\mathrm{G}}. \end{aligned}$$

From measurements made on the signal of figure 40 at a fixed gate voltage of 0·62 V, it was found that

$$\frac{p_{\theta 1}(1)}{p_{\lambda 1}(1)} = \exp\left[\frac{E_{\lambda 1}(1/0) - E_{\theta 1}(1/0)]}{kT}\right] = 1{\cdot}67,$$

and thus $E_{\lambda 1}(1/0) - E_{\theta 1}(1/0) \approx 12\,\mathrm{meV}$. In a similar fashion, it was found that $E_{\mathrm{F}} - E_{\theta 1}(1/0) \approx 10\,\mathrm{meV}$ and $E_{\mathrm{F}} - E_{\theta 2}(2/1) \approx 10\,\mathrm{meV}$. The position of the Fermi level was obtained from standard MOSFET analysis. The energy levels are shown in the inset to figure 42. The important point to note is that the occupancy levels $E_{\theta 1}(1/0)$ and $E_{\theta 2}(2/1)$ 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 hence as 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 $E(2/1)$ lies below $E(1/0)$. In such a system the occupancy of the defect would change directly from 0 to 2, missing out the charge-1 state. The extra electron–lattice interaction accompanying the capture of the second electron offsets the additional Coulombic interaction energy (Anderson 1975). Ngai and White (1981) have carried out model calculations of defects at the Si/SiO$_2$ interface and found both metastable and negative-U behaviour.

We have hitherto referred to the charge levels of the defect as 0, 1 and 2, and inspection of figure 40 shows that the amplitudes of the 0 → 1 and 1 → 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, owing to the roughly exponential dependence of carrier concentration on surface potential, $n^+/n^0 \approx n^0/n^-$. The local conductances obey a corresponding 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 $0 \rightarrow -$.

6.4. *Defect interactions in the oxide of MOS tunnel junctions*

Farmer *et al.* (1987, 1988) have recently studied MOS tunnel junctions. The devices under investigation were 1 μm^2 Al–SiO$_2$–p-Si diodes. The tunnelling current was due to electrons tunnelling between the metal and the conduction band of the silicon. Since the substrate of the device was p-type, the device is referred to as a minority-carrier tunnel diode. Green *et al.* (1974) and Shewchun *et al.* (1974) give excellent accounts of the basic device physics.

As might be expected, RTS behaviour was observed in the tunnelling current. The RTS signals can be conveniently classified as conventional and anomalous: figures 43 (*a*) and (*b*) provide examples of the former, and figures 43 (*c*) and (*d*) exemplify the latter. Figure 43 (*a*) shows the familiar two-level switching behaviour, and figure 43 (*b*) displays the effect of two RTSs that switch independently of one another. In figure 43 (*c*) a three-level signal is shown, which the authors say is not the result of summing two independent RTSs. In figure 43 (*d*) the small-amplitude jumps occur only from the lowest level.

Two aspects of the conventional RTS results that Farmer *et al.* (1987) found to be inconsistent with single-electron capture were the following. First, with $\tau = \tau_0 \exp(E/kT)$, values of the attempt frequency $1/\tau_0$ were found to reside in the range 10^2–10^{12} s^{-1}. They noted that the smaller values for $1/\tau_0$ were many orders of magnitude too low to be the rate for electron tunnelling through a 1·5–2·0 nm SiO$_2$ barrier. Second, values of $\Delta R/R$ were found ranging from less than 0·1% to greater than 10%. Using the model of Schmidlin (1966), one would expect a resistance change of about 0·01% for capture of a single electron in an otherwise uniform 1 μm^2 diode.

These two observations together with the anomalous RTS behaviour of figures 43 (*c*) and (*d*) led Farmer and co-workers to the conclusion that what they were observing was not independent fluctuations in occupancy of single defects but rather the fluctuations in charge state of strongly interacting defects. This hypothesis accounts for the low attempt frequencies (since co-operative capture of many hundreds of electrons is required), the large fractional change in resistance, and the three-level RTS signals, which are the result of trap–trap interactions.

Two further interesting features of the studies by Farmer *et al.* are the behaviour of some large-amplitude RTSs as a function of temperature and the temporal behaviour of the diode resistance at large applied bias. An example of the first effect is shown in figure 44. At the lowest temperature there is sharp two-level switching; and as the temperature is increased further, intermediate levels become apparent while the extrema still remain. They concluded that this was yet further evidence for the inter-defect interaction model. At low temperatures the two-level RTS is caused by ensembles of strongly interacting traps emptying and filling simultaneously. As the temperature is increased, this interactions weakens, so that only sub-ensembles of traps interact, thus accounting for the intermediate steps.

Figure 45 shows the time development of the diode resistance in the presence of an applied electric field of 18 MV cm^{-1}. It was found that the resistance was stable for long periods of time, but at random intervals a transition to a lower level took

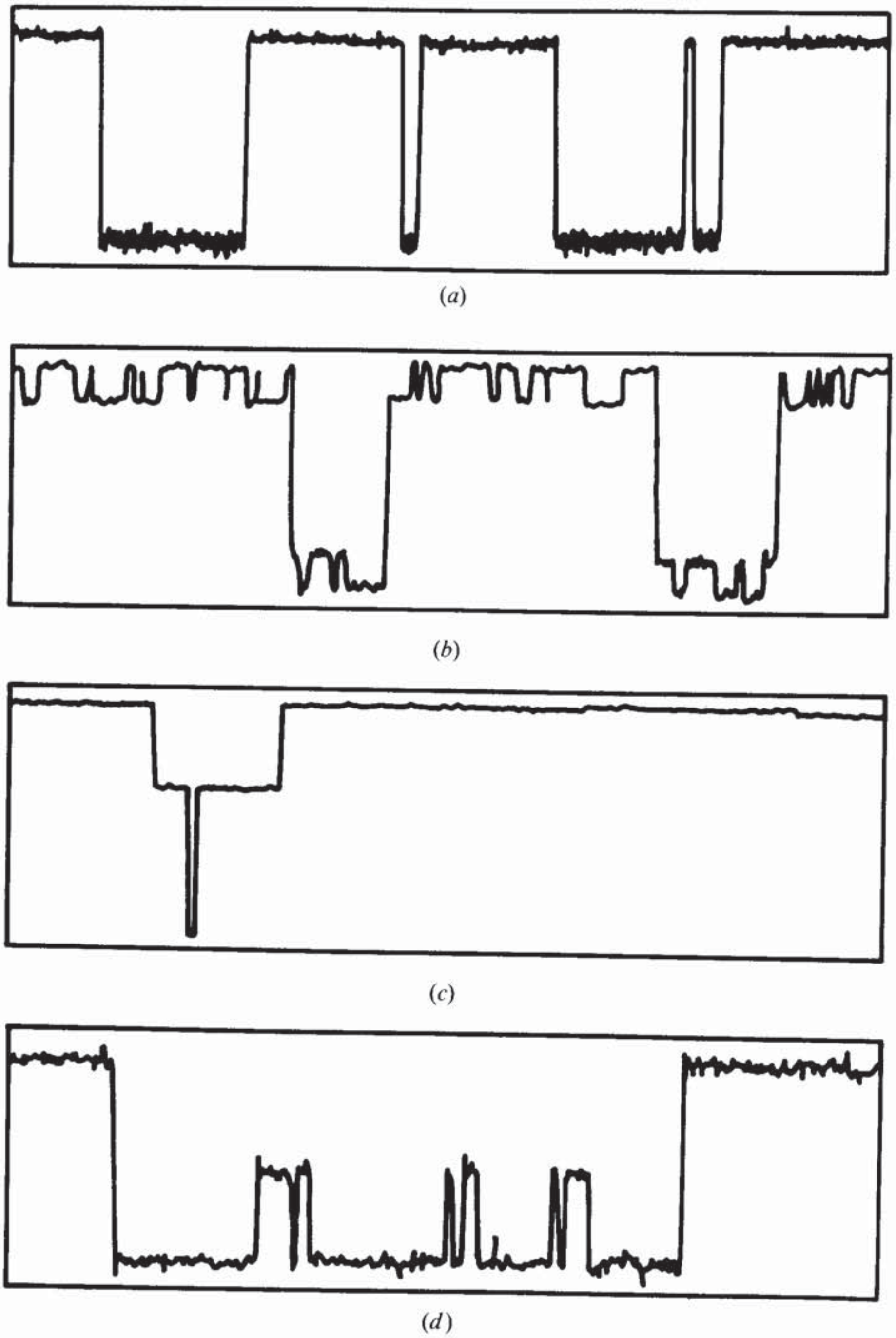


Figure 43. Time records of MOS tunnel-diode resistances showing some of the different types of discrete low-frequency fluctuations observed at moderate bias. (*a*) V = 1·41 V, T = 119 K; (*b*) 1·41 V, 77 K; (*c*) 1·41 V, 176 K; (*d*) 0·96 V, 194 K. From Farmer *et al.* (1987). © American Physical Society. Reproduced with permission.

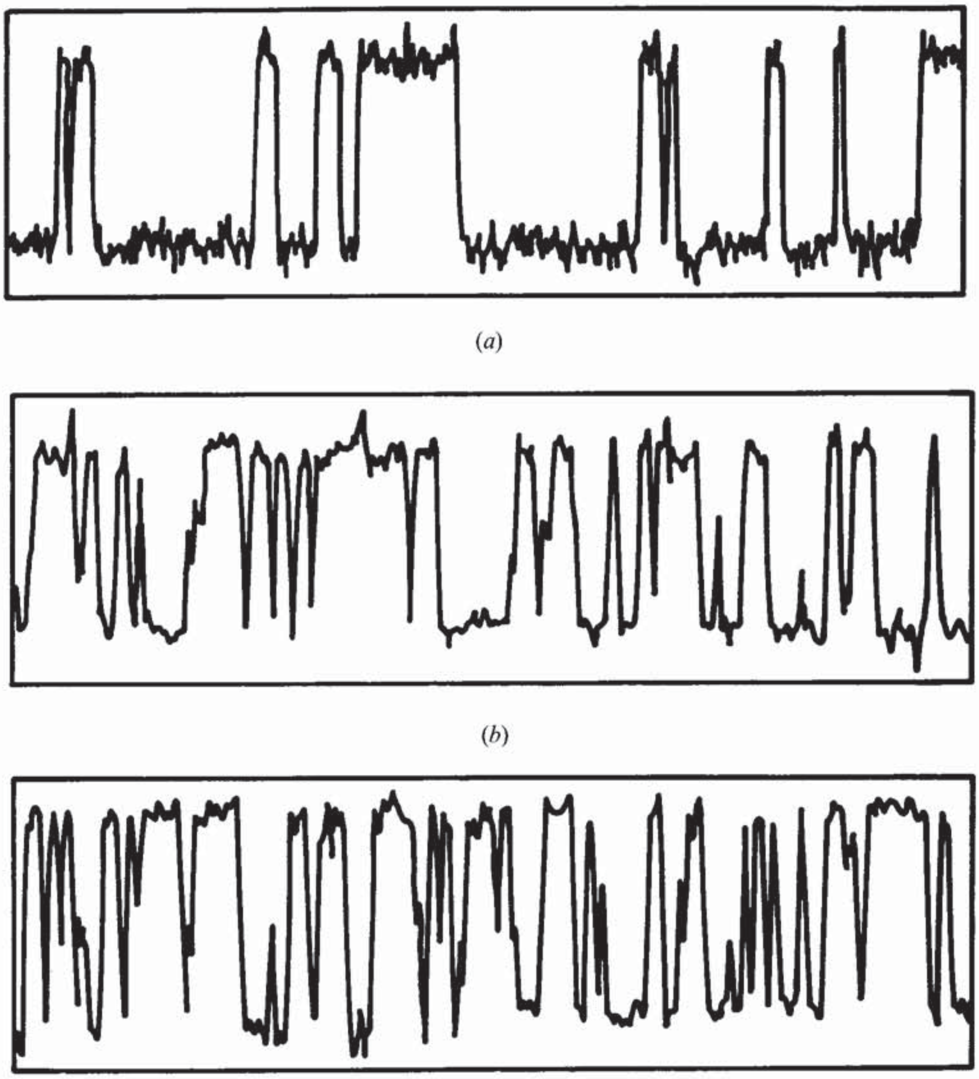


Figure 44. Resistance against time for a large RTS taken at three different temperatures ($V = 1{\cdot}50$ V). (*a*) At $T = 77$ K the RTS switches sharply between two very well defined levels. (*b*) At $T = 174$ K the RTS pauses at a few well defined intermediate values. (*c*) At $T = 226$ K the RTS switches between still more levels. After Farmer *et al.* (1987). © American Physical Society. Reproduced with permission.

place, followed by a complex sequence of switching events. They suggested that the 'breakdown' events in these figures were initiated by the collective action of groups of defect states, leading to irreversible changes in the structure of the SiO_2 film. This part of the work was further expanded upon in a subsequent publication (Farmer *et al.* 1988).

Several aspects of Farmer and co-workers' work are clearly worthy of further investigations. First of all, the homogeneity of the barrier will play a critical role in

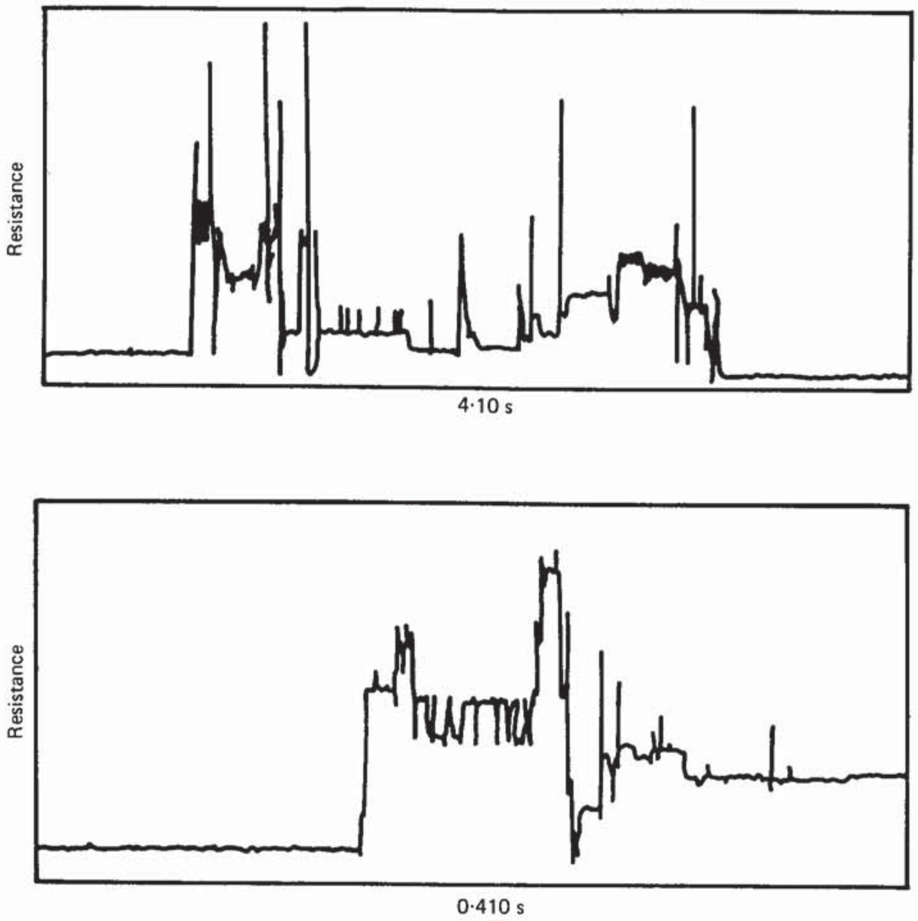


Figure 45. Resistance *versus* time for an MOS diode biased at high voltage ($V = -2{\cdot}99\,\mathrm{V} \equiv 18\,\mathrm{MV\,cm^{-1}}$), with $I = 0{\cdot}100\,\mathrm{mA}$. The upper half shows the record of a single breakdown event ($\delta R/R = 6{\cdot}7\%$ full scale). The resistance after the event is noticeably lower than before. The lower half, an expanded portion of the complete time trace, shows that the switching is discrete and quasi-stationary for brief periods of time. After Farmer *et al.* (1987). © American Physical Society. Reproduced with permission.

determining the resistance change. Preferential current flow through a small portion of the diode will give rise to a much larger $\Delta R/R$ than anticipated. Neri *et al.* (1987) have also investigated tunnelling currents through small-area MOS tunnel diodes and suggested that the presence of resonant tunnelling centres may also be important. Such a centre might add to the current flow through the device by permitting N electrons s^{-1}, say, to pass through it. The centre could be turned off by some localized change in its environment caused by single-electron capture. As we pointed out earlier (section 3.2.3), resistance changes greater than accounted for by one electron do not necessarily militate against single-electron capture. Indeed, one of the outstanding problems in this field is to understand the modulation in resistivity brought about by trapping one electron. Finally, while the strongly interacting, multi-electron trap model would seem to superficially explain the data, much more work needs to be done on the detailed kinetics of the model. In this regard, further data as a function of bias and temperature are required.

6.5. *Noise in metallic microstructures*

Metal films exhibit low-frequency current fluctuations when biased with a constant voltage. The power spectrum scales approximately as $1/f$, and the noise arises from intrinsic resistance fluctuations. There is now considerable evidence linking defects with the noise production: see for example Dutta and Horn (1981), Weissman (1988), Koch *et al.* (1985), Pelz and Clarke (1985) and Fleetwood and Giordano (1985). Dutta and Horn (1981) showed that the noise could be explained by a thermally activated process with a wide distribution of activation rates and suggested that defects might be responsible. Koch *et al.* (1985) investigated the $1/f$ noise in polycrystalline films of aluminium. The temperature dependence of the noise indicated activation energies similar to those found for Al diffusion along the grain boundaries.

Pelz and Clarke (1985) carried out a particularly nice study of the noise in copper films under 500 keV electron irradiation at 90 K. They found that the noise level increased under irradiation much more than the sample resistivity, and that annealing of the noise and resistivity took place in different temperature regimes. This work provided strong evidence that the defects responsible for the $1/f$ noise were distinct from those responsible for the residual resistivity.

Despite the overwhelming evidence linking defects to noise production in metals, the mechanisms that couple the defect to the modulation in resistivity still remain uncertain. As pointed out by Weissman (1988), the carrier-trapping or number-fluctuation models are not appropriate since the required number of fluctuating traps would be very large. Thus one must look for fluctuations in carrier mobility.

On the basis of the work of Martin (1971, 1972) and Kogan and Nagaev (1982), Pelz and Clarke (1987) have recently calculated the resistance fluctuations due to changes in scattering cross-sections as defects move (rotate). They denoted their model a 'local-interference' model, since the effect is predominant in defects composed of multiple scattering centres within one or two lattice constants of one another. Such centres might be, for instance, first- and second-neighbour divacancies, trivacancies, split interstitials and suchlike. It is to be noted that monovacancies and substitutional impurities would not show the necessary marked anisotropy of the cross-section.

Discrete switching in the resistance of metallic nanobridges has very recently been observed by Ralls and Buhrman (1988): see figures 46 (*a*)–(*d*). The structures under investigation were clean copper constrictions of sizes 40–8000 nm^3. Figure 46 (*a*) shows the conventional two-level switching. Whereas in the MOSFET case the switching is associated with carrier capture and emission events at single defects, here the switching is due to the defect transforming from one minimum-total-energy configuration to another. As it does so, its scattering cross-section changes, resulting in a modulation of the sample resistance. Figure 46 (*b*) shows the superposition of two independent RTSs.

Particularly interesting behaviour is depicted in figures 46 (*c*) and (*d*): figure 46 (*c*) shows that the amplitude of the small RTS is larger when the large RTS is in its down state; figure 46 (*d*) shows that the rapid fluctuations completely disappear for extended periods. In this respect, these signals bear a strong resemblence to the complex telegraph noise observed in small MOSFETs, which we discussed at length in sections 6.2 and 6.3. We concluded that the complex telegraph noise in MOSFETs

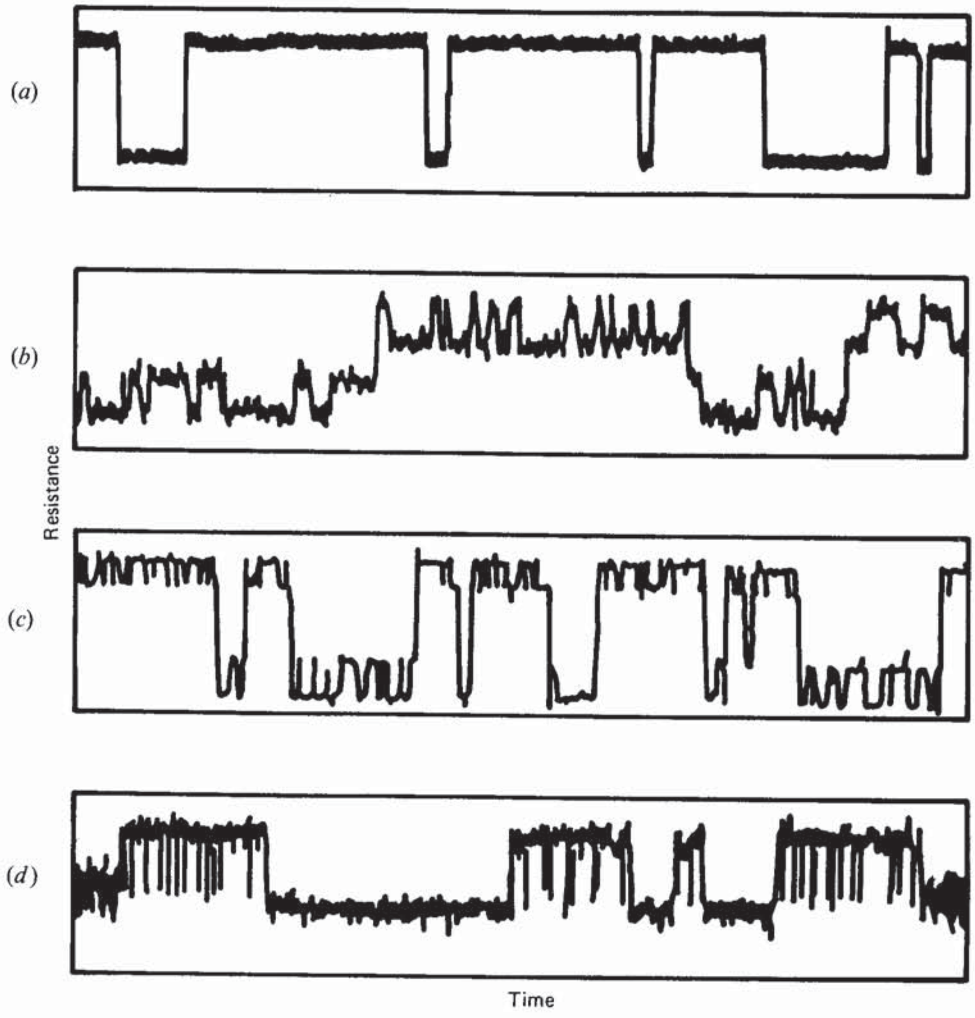


Figure 46. Resistance against time in copper nanobridges for $T < 150$ K, showing several types of behaviour (see the text). The fluctuations studied range from 0·005% to 0·2%. From Ralls and Buhrman (1988). © American Physical Society. Reproduced with permission.

was probably due to individual defects exhibiting metastability. Now the work of Pelz and Clarke indicates that in metals complex defects are most likely to give rise to resistance modulations. For a complex defect, one can envisage the existence of a number of metastable total-energy minima in some general configuration space. The scattering cross-section will be different for each configuration. In figure 46 (*c*) the small-amplitude fluctuations at the top and bottom of the trace correspond to different overall defect configurations; thus it is not necessarily surprising that the amplitudes differ. Similarly, for the trace in figure 46 (*d*) the extended periods in which the fluctuations cease might well correspond to the complex defect having transformed into a very different metastable minimum. These ideas are compatible with the discussion of metastability in section 6.2.

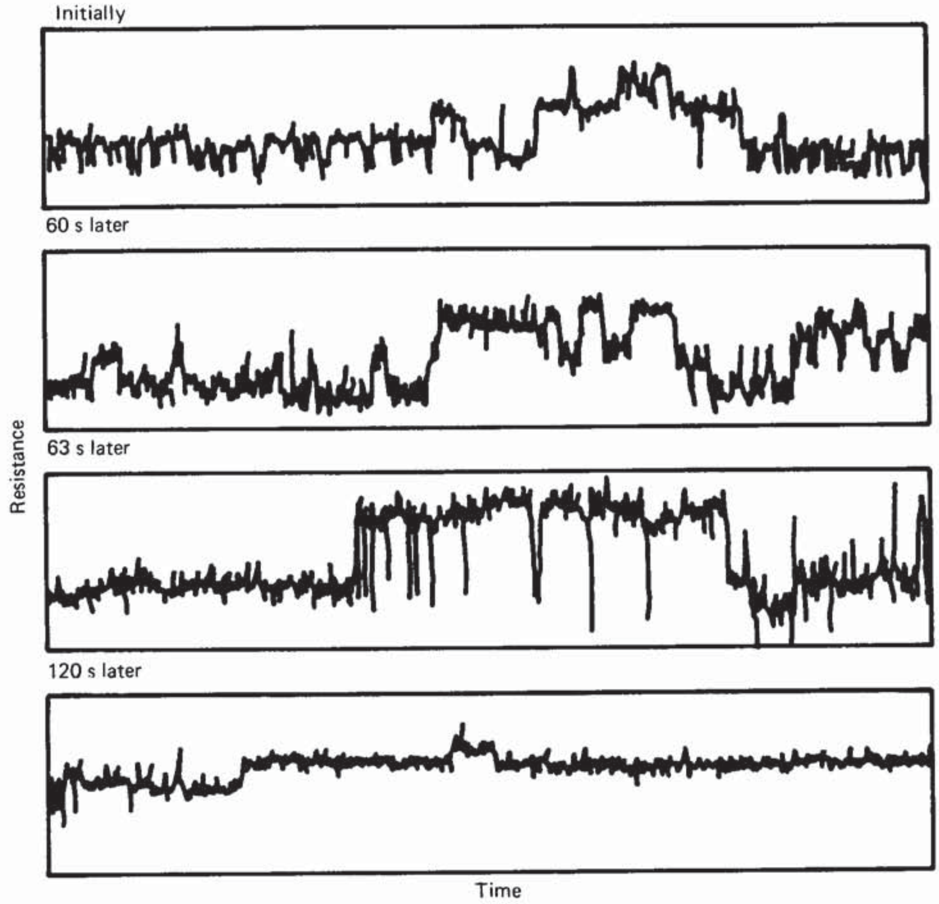


Figure 47. Resistance against time for a 90 Ω nanobridge at T = 300 K, displaying the wandering nature of the low-frequency noise, which is still composed of discrete resistance fluctuations. Each noise snapshot is about 0·2 s long. From Ralls and Buhrman (1988). © American Physical Society. Reproduced with permission.

Further evidence for the existence of low-lying, metastable, total-energy minima comes from the traces in figure 47. This figure shows 0·2 s snapshots of the switching events, with each trace separated by about 60 s. It is evident that the switching evolves in a very complex manner. Indeed, Ralls and Buhrman found the trace to exhibit a $1/f$ spectrum when averaged over several minutes.

6.6. *Complex telegraph noise in MOSFETs (III): Switching without change of charge state*

In section 3.2.3 we addressed the problem of the rather wide distribution of RTS amplitudes (figure 15), and suggested that some additional mechanism beyond number fluctuations was responsible for the RTS step height. In section 6.5 we saw that in metals a defect can modulate the resistivity through a geometric transformation that brings about a change in scattering cross-section. The question of whether a defect transformation without any change in charge state can modulate the channel conductance in a MOSFET is what we wish to consider here. If such a mechanism

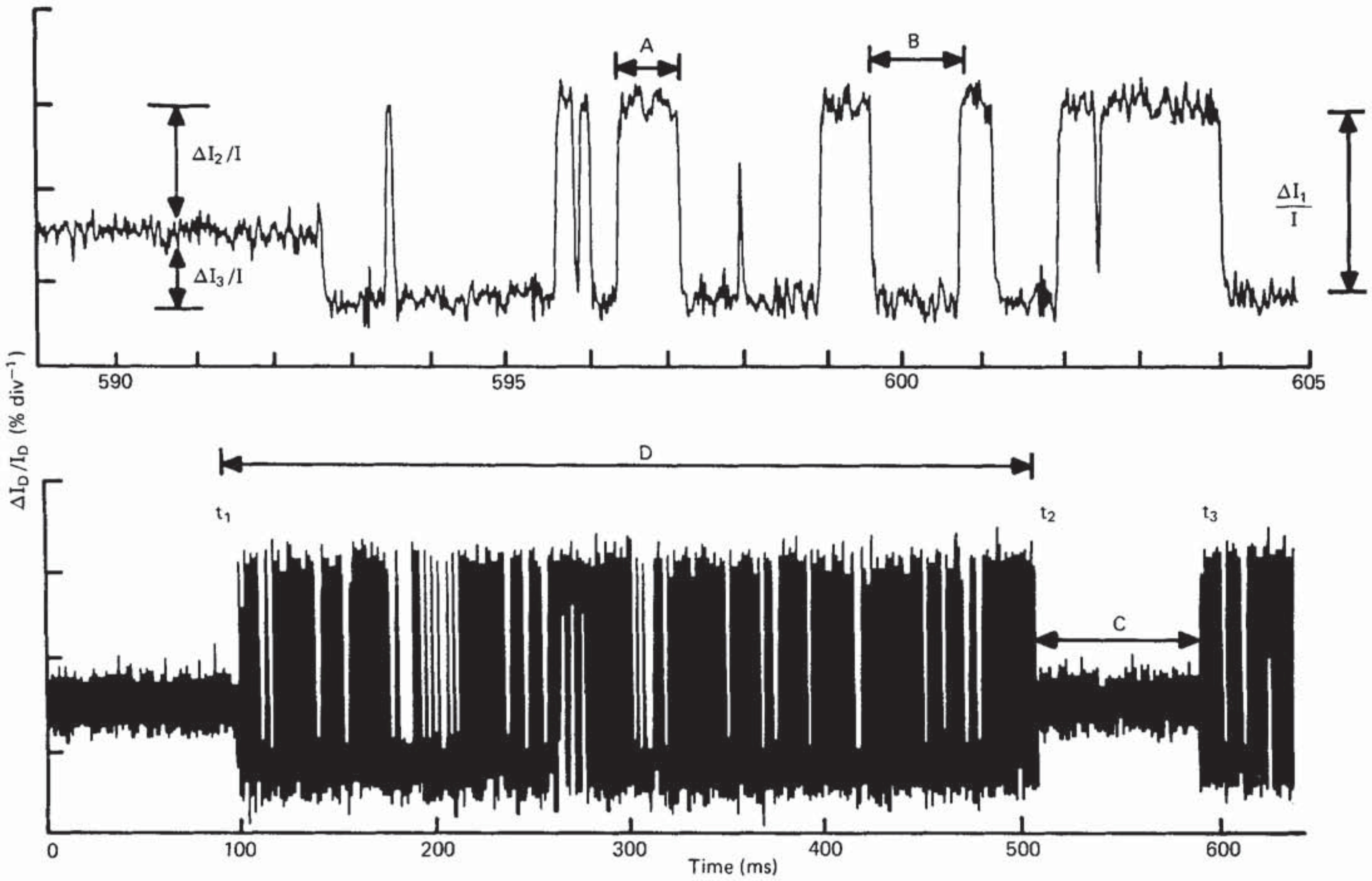


Figure 48. Fluctuations in current against time, showing a rapidly switching RTS modulated by an envelope of different amplitude in a small (0·7 μm × 0·5 μm) MOSFET. The upper trace is an expansion of part of the lower trace. $V_G = 0{\cdot}55\,\mathrm{V}$, $V_D = 20\,\mathrm{mV}$, $I_D = 18{\cdot}8\,\mathrm{nA}$ and $T = 293\,\mathrm{K}$.

could be demonstrated then it would go some way towards providing an explanation for the widths of the amplitude data of figure 15.

Figure 48 shows the complex switching measured in a small (0·7 μm × 0·5 μm) MOSFET supplied to us by Dr O. Jantsch of Siemens Research Centre Munich. In outline, the RTS is similar to the anomalous RTS of figure 38: the underlying rapidly switching RTS is modulated by an outer envelope. We shall refer to the mean up time of the envelope as $\bar{D}$ and the mean down time as $\bar{C}$. When the fluctuations cease, as in the period from t_2 to t_3, the current level lies above the baseline; it is thus a three-level signal.

Figure 49 shows the dependence of $\bar{A}$ and $\bar{B}$ as functions of channel conductivity. Clearly $\bar{A}$ corresponds to carrier capture and $\bar{B}$ to carrier emission. We described the process for calculating the distance of the defect into the oxide in section 3.2.2. In addition, we showed in section 5.1.4 that the most reliable results are obtained when the device is operating away from threshold. Assuming single-electron trapping and using equation (3.4) with the sub-threshold data points of figure 49, we find that for a δV_G of 25 mV, $\delta\phi_s = 18$ mV and $\delta(\Delta E_{TF}) = 19{\cdot}6$ meV. With a gate-oxide thickness of 18 nm, this places one trap about 4 nm from the interface. On the other hand if we assume two-electron trapping (equation (3.5)) then we find $\delta(\Delta E'_{TF}) = 9{\cdot}8$ meV, which would place the defect 85 nm into the bulk silicon. In addition, the defect would show very strong negative-U properties ($U_{eff} < -0{\cdot}4$ eV) and have a cross-section for capture of the second electron of about 10^{-13} cm^2. While two-electron trapping clearly cannot be ruled out, single-electron trapping offers the simplest model for the rapid switching shown in figure 48.

The behaviour of the time $\bar{C}$ with channel conductivity is shown in figure 50. Since $\bar{C}$ increases with increasing inversion-layer number density, we can conclude that it

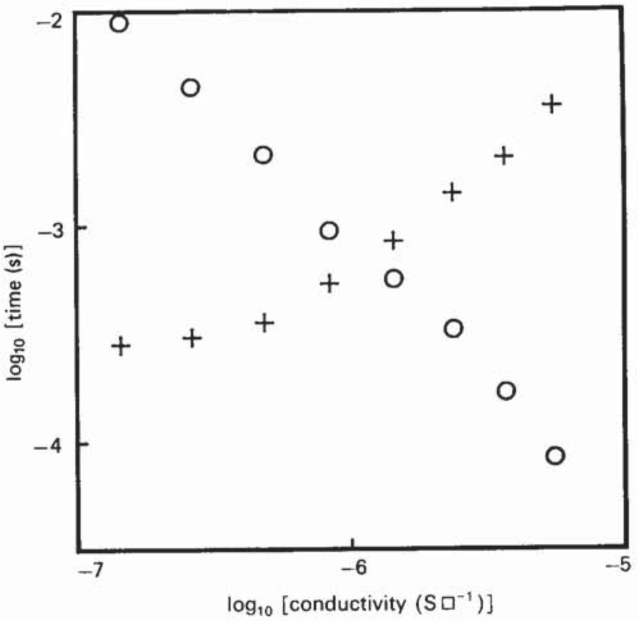


Figure 49. The dependence of the times $\bar{A}$(○) and $\bar{B}$(+) on channel conductivity for the RTS of figure 48.

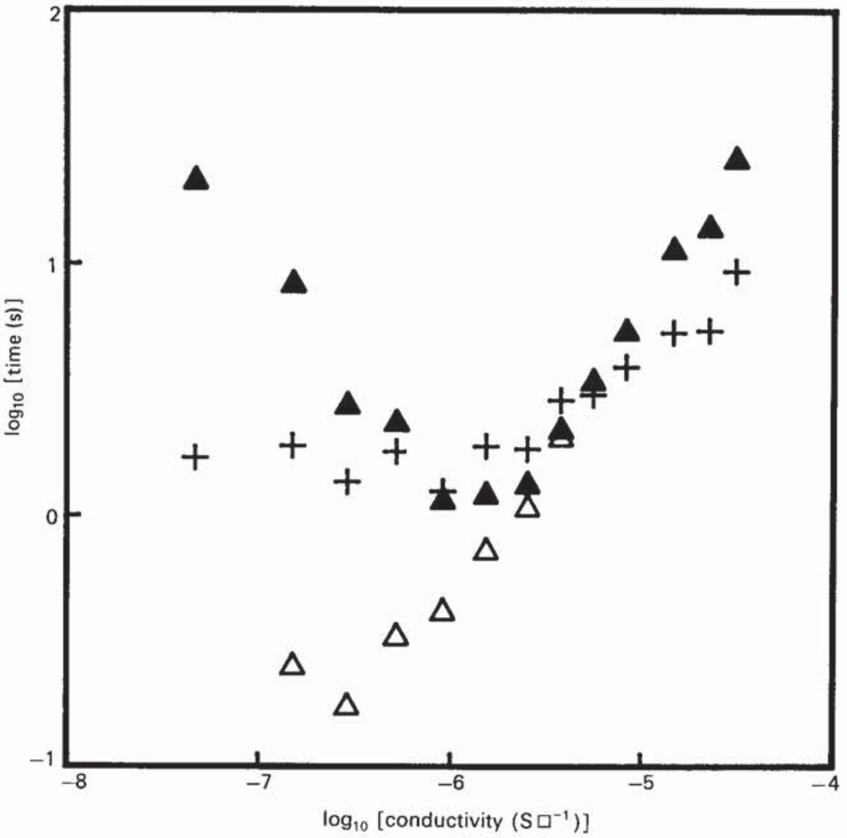


Figure 50. The dependence of times $\overline{D}$(▲), $\overline{C}$(+) and $\overline{\Sigma B}$(△) on channel conductivity ($= I_D l/V_D w$) for the RTS of figure 48.

does not correspond to carrier capture. In the interacting-traps model (model (*a*) of section 6.2) *C* would correspond to the time taken for the slow trap to emit its captured electron. This would be inconsistent with the fact that the current always decreases at the end of time period *C*. In none of our measurements on RTSs at room temperature have we ever observed the current to decrease on electron emission. Of course one could argue that at the beginning of the time period from t_2 to t_3 electron capture from the high-current level is always so fast that it appears on the trace as though the level *C* is approached from the lowest level. Similarly, at the end of the period *C* the current increases but capture is so fast that it looks as though the current level goes straight down. The problem with this argument is simply the fact that if capture into the middle level were extremely rapid then one would never observe the fluctuations *A* and *B*. Thus the two-trap model, or its equivalent, metastable-state model (*a*), cannot provide a straightforward explanation of the data.

An appealing interpretation of the data is the following. The trace shown in figure 48 is the result of a single defect. The times *A* and *B* correspond to single-electron charging and discharging of the defect in total-energy minimum α_1. In addition, the defect possesses a metastable state β_1, into which it transforms and resides for extended periods of time *C*—see the configuration-coordinate diagram of figure 39 (*b*). Since the model of Dierickx (section 6.2) is equivalent to the model of figure 39 (*b*), one cannot differentiate between his model and model (*b*). The key difference between this defect and the defects described in section 6.2, however, is that on transformation, with no change in charge state, there is a change in channel conductivity and the current increases to the level shown in the period from t_2 to t_3.

The mean time taken for the defect to transform from α_1 to β_1 is given by the sum of the down times over an average time period from t_1 to t_2. To be consistent with our previous notation, we shall refer to this time as $\overline{\Sigma B}$. The dependence of $\overline{\Sigma B}$ and $\overline{D}$ on channel conductivity is shown in figure 50. For a given current level, the time $\overline{\Sigma B}$ has been obtained from $\overline{D}$ by multiplying the latter by the fraction of the time the rapidly switching RTS is in its down state (i.e. $\overline{B}/(\overline{A} + \overline{B})$). It is evident that the two transformation times $\overline{C}$ and $\overline{\Sigma B}$ are relatively constant below threshold, and then start to rise. In the RTS of figure 38 we found the transformation times to be independent of gate voltage. However, here there appears to be a strong interaction between the trapped charge in its transformed state and the inversion layer. So it is not surprising that the transformation times show a dependence on inversion-layer number density. We might expect $\overline{C}$ and $\overline{\Sigma B}$ to show the same V_G dependence; it is similar, but not identical.

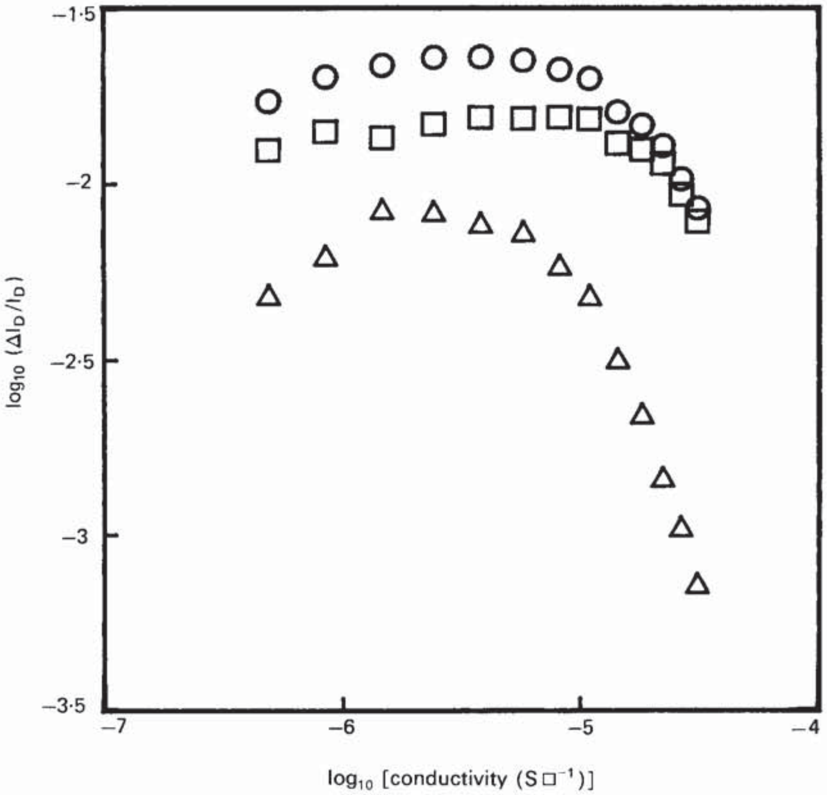


Figure 51. The behaviour of amplitudes $\Delta I_1/I$(○), $\Delta I_2/I$(□) and $\Delta I_3/I$(△) with channel conductivity.

The dependence of the various step heights of the RTS on channel conductivity is depicted in figure 51. The origin of the slight decrease in the magnitude of $\Delta I_1/I$ at low conductivities is not clear, but may have something to do with the lightly-doped-drain device topography. However, the rate of decrease of $\Delta I_1/I$ in strong inversion is that expected for the screening of one trapped electron, see section 3.2.3. Evidence for the different physical origin of $\Delta I_3/I$ is provided by the way in which this amplitude decreases very rapidly with increasing inversion-layer number density. As to the physical mechanism underlying $\Delta I_3/I$, one can only speculate. Presumably, on defect transformation a different lattice polarization will be set up, and this effect may be enhanced by the proximity of any fixed oxide charge. In addition, the polarization may affect the interface phonons. For Dierickx's model the slower trap is further into the oxide and this may account for its smaller amplitude. While it is always dangerous to argue from the specific example to the general case, the RTS of figure 48 provides evidence that the RTS step height may depend critically on other (as yet unknown) factors in addition to the number-density changes brought about by purely Coulombic interactions. Clearly there is much scope for further fruitful investigation in this area.

7. The Si/SiO$_2$ interface

The single most important interface in semiconductor technology is that between thermally grown silicon dioxide and the underlying silicon. Much of what we have discussed up to this point has been concerned with the rather specific additional insights into the nature of trapping centres associated with this interface that have come from the study of MOSFET microstructures. In the light of this new information, we wish to broaden our enquiry. Our goal will be to provide a unified description and understanding of the results obtained from two of the major electrical characterization techniques that have been applied to the oxide/silicon interface: these are noise and conductance (G–ω) measurements.

In section 4 we noted that in condensed-matter microstructures $1/f$ noise decomposes into its individual fluctuating components; sections 5 and 6 then addressed the problem of understanding the dynamics of these fluctuations. On the basis of the measured properties of individual defects, in section 7.1 we shall consider in detail the generation of $1/f$ noise in MOSFETs from a summation of these individual noise-generating events.

Capacitance–voltage (C–V) and conductance measurements are the everyday experimental techniques used to assess the quality of the oxide/silicon interface in MOS structures. The presence of interface states gives rise to the familiar phenomena of stretch-out in the high frequency C–V curve and the loss peak in G–ω data (Nicollian and Brews 1982, Schulz 1983). Conventional wisdom has it that at a given energy in the silicon band gap the interface states have only a single well defined cross-section. In sections 7.2 and 7.3 we show that in device-quality oxides the defect states in the oxide (which reside close to the interface and are responsible for random telegraph signals and $1/f$ noise) are observable in conductance measurements at room temperature. A loss peak due to the normal interface states is seen, together with a plateau on the low-frequency side due to these trapping centres in the oxide.

7.1. *1/f noise in MOSFETs*

Here we have two very specific aims: first, to search for detailed agreement between numerical simulations and experiment for the case of the current spectral density in silicon MOSFETs; and second, to investigate the physics underlying the wide distribution of (trapping) time constants responsible for producing $1/f$ noise. To treat the first problem, we shall initially assume that such a distribution of time constants does exist.

7.1.1. *The noise current spectral density: numerical and experimental results*

Using equation (2.19 *a*), we can write the current spectral density in a MOSFET in the form

$$\frac{S_I(f)}{I^2} = \sum_{k=1}^{N_{\rm dev}} \frac{4(\Delta I/I)_k^2}{(\bar{\tau}_{\rm c} + \bar{\tau}_{\rm e})_k[(1/\bar{\tau}_{\rm c} + 1/\bar{\tau}_{\rm e})_k^2 + (2\pi f)^2]}, \tag{7.1}$$

where the index k runs over all RTSs, $N_{\rm dev}$, in the device of area A, and k labels the fractional amplitude and time constants associated with the kth RTS. We assume that the defects are uniformly distributed on the interfacial plane. Our intention is to use the data obtained from small-area devices of area a ($\approx 0{\cdot}4\,\mu{\rm m}^2$) to generate power spectra for a device of any size. We have already shown in section 4.1 that

$$\frac{\Delta I}{I} \propto \frac{1}{A}.$$

Thus if in a small device of area a the measured value of $\Delta I/I$ is equal to $\Delta I_{\rm m}/I_{\rm m}$ then in a device of area A

$$\frac{\Delta I}{I} = \frac{a\,\Delta I_{\rm m}}{AI_{\rm m}.} \tag{7.2}$$

Similarly the average number of RTSs in the device of area A, $N_{\rm dev}$, can be evaluated by counting the number of RTSs visible in a given time window, $n_{\rm w}$, at a given gate voltage in small device: $N_{\rm dev} = (A/a)n_{\rm w}$. Substituting the above information into equation (7.1), we find that the normalized spectral density is given by

$$\frac{AS_I(f)}{I^2} = \frac{1}{A}\sum_{k=1}^{An_{\rm w}/a} \frac{4(a\,\Delta I_{\rm m}/I_{\rm m})_k^2}{(\bar{\tau}_{\rm c} + \bar{\tau}_{\rm e})_k[(1/\bar{\tau}_{\rm c} + 1/\bar{\tau}_{\rm c})_k^2 + (2\pi f)^2]}. \tag{7.3}$$

(It is clear that the spectrum is normalized since the upper limit on the summation is directly proportional to the device area A, while the denominator contains the term A explicitly.)

Figures 21 (*a*)–(*c*) depict the normalized power spectra measured in silicon MOSFETs with active device areas of 350, 15 and $0{\cdot}4\,\mu{\rm m}^2$ respectively. As expected, there is significant sample-to-sample variation in the smallest devices; but the average spectrum resides at a level compatible with those of the large-area devices. For us to use equation (7.3) to generate numerically spectra equivalent to those of figures 21 (*a*)–(*c*), two further pieces of information are required: the distribution of amplitudes, $\Delta I_{\rm m}/I_{\rm m}$, measured for RTSs in small-area devices; and the average number of RTSs, $n_{\rm w}$, visible within a given time window and under given operating conditions. With respect to the first of these requirements, on figure 15 we show a percentile plot of the measured fractional amplitudes against channel conductivity. The data at $G = 5{\cdot}36 \times 10^{-5}\,{\rm S}\,\square^{-1}$ were obtained from 56 RTSs and correspond to $V_{\rm G} \approx 2{\cdot}2\,{\rm V}$. We do not wish to comment again on the nature of the distribution of trap amplitudes at this $V_{\rm G}$ value beyond noting that, on a logarithmic scale, it is approximately normally distributed with mean $-3{\cdot}0$ and standard deviation $0{\cdot}2$.

In order to obtain a $1/f$ spectrum, the time constants need to be distributed uniformly on a log scale (Dutta and Horn 1981). Hence the number of traps (per unit time interval) with time constants residing in the interval from τ to $\tau + d\tau$ is given by

$$n(\tau) = c/\tau, \tag{7.4}$$

where c is a constant. The total number of RTSs with mean time constants between t_1 and t_2 is given by

$$N(t_1, t_2) = \int_{t_1}^{t_2} n(\tau)\, d\tau = c \ln\left(\frac{t_2}{t_1}\right).$$

The mean time constant τ of an RTS is chosen from a uniform distribution on a logarithmic time scale so that $\tau = 10^p$, where p is a random number in the interval $P_1 \leqslant p \leqslant P_2$. An RTS is visible only if its mark-space ratio is close to unity. The mark-space ratio is set by the separation of the trap energy level and the Fermi level via the relation $\bar{\tau}_e = \bar{\tau}_c \exp[-(E_T - E_F)/kT]$. We assume that the energies E_T are uniformly distributed and take $\bar{\tau}_e = \bar{\tau}_c \exp(q)$, where q is a random number chosen such that $-Q \leqslant q \leqslant Q$. A cut-off of $Q = 2{\cdot}0$ is used, thus ensuring a mark-space ratio close to unity. Combining the above analysis with the relationship

$$\frac{1}{\tau} = \frac{1}{\bar{\tau}_c} + \frac{1}{\bar{\tau}_e}, \tag{7.5}$$

we find

$$\bar{\tau}_c = 10^p[1 + \exp(-q)], \tag{7.6a}$$

$$\bar{\tau}_e = 10^p[1 + \exp(q)]. \tag{7.6b}$$

Note that the equations (7.3) and (7.6 *a*, *b*) are symmetric with respect to interchange of $\bar{\tau}_c$ and $\bar{\tau}_e$.

The results of numerical simulations of the current spectral density in devices of area 350, 15 and 0·4 μm^2 are given in figures 52 (*a*)–(*c*). In the generation of these figures, it was assumed that there was one RTS per decade in time. This was based on the experimental observation that in a 0·4 μm^2 device there were, on average, about four RTSs active in a time window stretching from 10^{-3} to 10 s. The values of τ were chosen from a uniform logarithmic distribution between 10^{-5} and 10 s. Comparing figures 52 (*a*) and 21 (*a*), we find very good overall agreement, over the full frequency range, between the numerical simulations and the experimentally measured power spectra. Similarly, the sample-to-sample variation is evident in the simulated spectrum of figure 52 (*c*). One feature that is slightly surprising is the larger than anticipated variation in the measured spectra for the 15 μm^2 devices (20 μm × 2 μm as drawn). This is perhaps due to inhomogeneities in channel length from device to device.

We have also investigated the sensitivity of the above numerical results to the variation of the various parameters involved. In particular, variations in the number of RTSs visible per decade in the 0·4 μm^2 devices and the allowed range for the aspect ratio. The effects of these changes are simply multiplicative and not dramatic. Nevertheless, the foregoing figures demonstrate quite convincingly that a reasonably accurate description of the noise properties of large-area silicon MOSFETs can be predicted from the individual fluctuations measured in small devices.

A convenient measure of the noise properties of a MOSFET is the observed power spectral density at a given frequency plotted as a function of the device's conductivity. In figure 53 we show $AS_I(f = 2\,\text{Hz})/I^2$ measured in a 350 μm^2 device plotted against

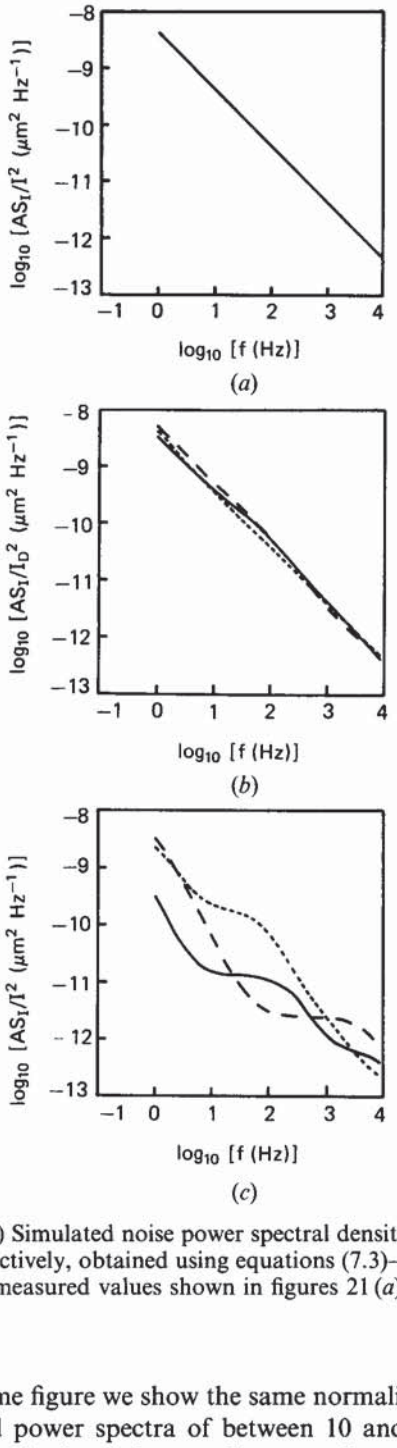


Figure 52. (*a*), (*b*) and (*c*) Simulated noise power spectral density in MOSFETs of area 350, 15 and 0·4 μm^2 respectively, obtained using equations (7.3)–(7.6). The spectra should be compared with the measured values shown in figures 21 (*a*)–(*c*).

conductivity. On the same figure we show the same normalized function obtained by averaging the measured power spectra of between 10 and 15 small-area (0·4 μm^2) devices. It is evident from the figure that excellent overall agreement between the small and large devices is obtained over a wide range of operating conditions. More importantly, it provides evidence that the measured distribution of amplitudes in the small devices is not being grossly distorted by any inhomogeneities or by the proximity

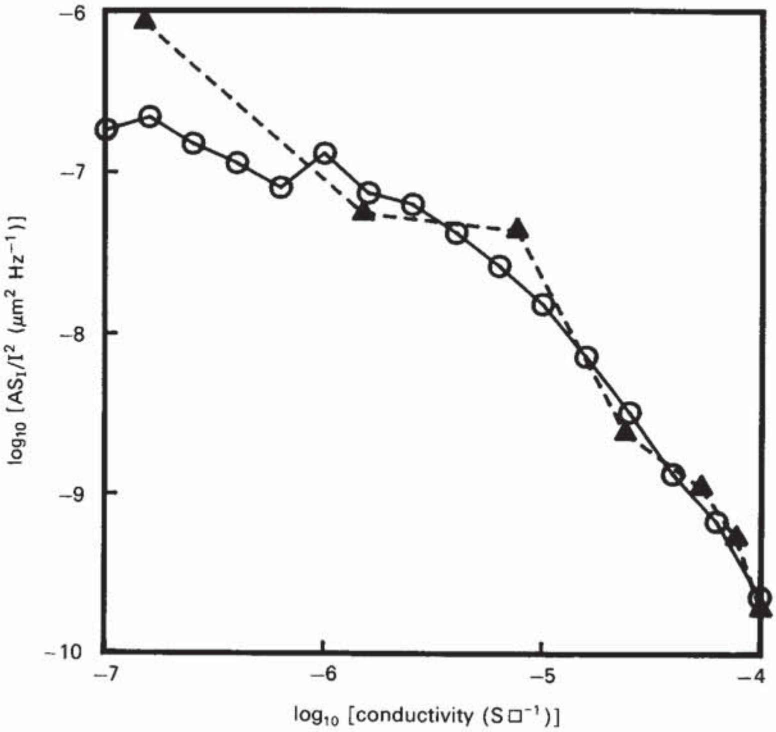


Figure 53. Noise power spectral density at $f = 2$ Hz against channel conductivity: ○ depicts normalized $S(f = 2\,\text{Hz})$ measured in a 20 µm × 20 µm MOSFET; ▲ depicts normalized $S(f = 2\,\text{Hz})$ obtained from averaging over measurements on 2 µm × 2 µm devices.

of the device perimeter. For example if the defects in the small devices were associated predominantly with the perimeter then, although the amplitudes of the RTSs would be smaller than expected (see section 3.2.3), the projected noise magnitude of the large devices would be severely overestimated.

7.1.2. *Distribution of trapping times*

The question that we now wish to address is the following: What physical mechanism is responsible for generating the particular distribution of time constants implicated in $1/f$ noise? With respect to the silicon MOSFET, we have seen that carrier capture into oxide traps close to the Si/SiO$_2$ interface states proceeds via a thermally activated multiphonon process. We shall now investigate the implications of this model for $1/f$ noise.

We can focus attention on the capture time $\bar{\tau}_c$, since the emission time $\bar{\tau}_e$ is given by the requirements of detailed balance. In section 5.1.2 we showed how the capture time is dependent upon the various device parameters. Equation (5.12) provided the key relationship

$$\bar{\tau}_c = \frac{\exp(\Delta E_B/kT)\; q\mu_0 V_D t(w/l)}{\sigma_0 (8kT/\pi m^*) I T^{3/2}} \qquad (7.7)$$

where the terms in this equation have exactly the same meaning as previously and we have substituted $t_0 = t/T$. So, for a device under given operating conditions, the distribution of capture times is generated via the permutation of σ_0 and ΔE_B among their many possible values. In this respect, the present limited data sets on ΔE_B and σ_0 (table 3) prove to be something of a handicap. However, on the basis of the data available, we shall start by making the following informed guesses for the distributions of ΔE_B and σ_0: ΔE_B varies uniformly between values E_1 and E_2 and $\ln \sigma_0$ varies uniformly between S_1 and S_2. Intuitively, both of these assumptions can be seen to be perfectly reasonable. Since the trapping centres reside in an amorphous material with its consequent almost-continuous distribution of trap environments, one expects a wide and smooth distribution of activation energies; the larger values corresponding to rotation and possible breaking of bonds. In the absence of any evidence to the contrary, it seems likely that the defects will be distributed evenly into the oxide. Our discussion in section 5.1.3 assumed that σ_0 depended upon the overlap of the initial- and final-state wavefunctions: thus one expects $\sigma_0 \propto \exp(-x/x_0)$, where x_0 is the decay length. Therefore we can write $\ln \sigma_0 \propto x$, leading to the values of σ_0 being uniformly distributed on a logarithmic scale.

Using values appropriate to the device whose power spectrum is depicted in figure 21 (*a*) we show in figures 54 (*a*)–(*d*) the distributions of capture times arising for a fixed range of ΔE_B and varying ranges of σ_0. In order to understand the form of these distributions, we shall rewrite equation (7.7) as

$$\begin{aligned}\log \tau_c &= \frac{\Delta E_B}{2{\cdot}303kT} - \log \sigma_0 + C \\ &\equiv A - B + C, \qquad (7.8)\end{aligned}$$

where C is a constant. The distribution of $\log \tau_c$ arises out of the convolution of the sets of numbers $\{A\}$ and $\{B\}$. The general result is that the flat portion of the final distribution is of size $\dim\{A\} - \dim\{B\} + 1$. For the example shown in figure 54 (*a*) the distribution $\{A\}$ extends over 17·4 (=40/2·303) decades and $\{B\}$ over one decade; $\log \tau_c$ is flat for just over 17 decades. As the distribution for σ_0 widens the width of the flat part of $\log \tau_c$ decreases as shown in figure 54 (*b*). When the activation energies and the pre-factors both extend over about 17 decades, the final distribution is triangular (figure 54 (*c*)). Thereafter, as the width of $\log \sigma_0$ increases, the height of $\log \tau_c$ decreases, but the width of its flat portion increases as shown in figure 54 (*d*).

Within the experimental time window $10^{-3}\,\mathrm{s} \leqslant \tau \leqslant 10\,\mathrm{s}$, the data of table 3 show that ΔE_B lies in the range 0·2–0·6 eV and σ_0 in the range 10^{-19}–$10^{-15}\,\mathrm{cm}^2$. Now if both ΔE_B and σ_0 were both very wide and uniform distributions and there were no correlation between the two, we would expect this to be reflected in the measured data. For example, a small value of ΔE_B would be associated with a small value of σ_0 such that τ resided in the time window. If one of the distributions were much narrower than the other then it would automatically constrain the measured width of the other distribution. This appears to be the experimental situation since the ranges are found to be similar: $\log \sigma_0$ spans about four decades and ΔE_B about seven decades. It is not possible experimentally to determine which has the narrower distribution. However, it is very easy to envisage σ_0 extending over many decades. So it is possible that σ_0 has the wider distribution. In practice, we shall see that the precise details of the distributions of σ_0 and ΔE_B hardly affect the noise power spectrum.

We shall taken ΔE_B to vary uniformly between 0 and 1 eV; $\log \sigma_0$ will be assumed to vary uniformly between some variable lower limit and an upper bound of -15. The

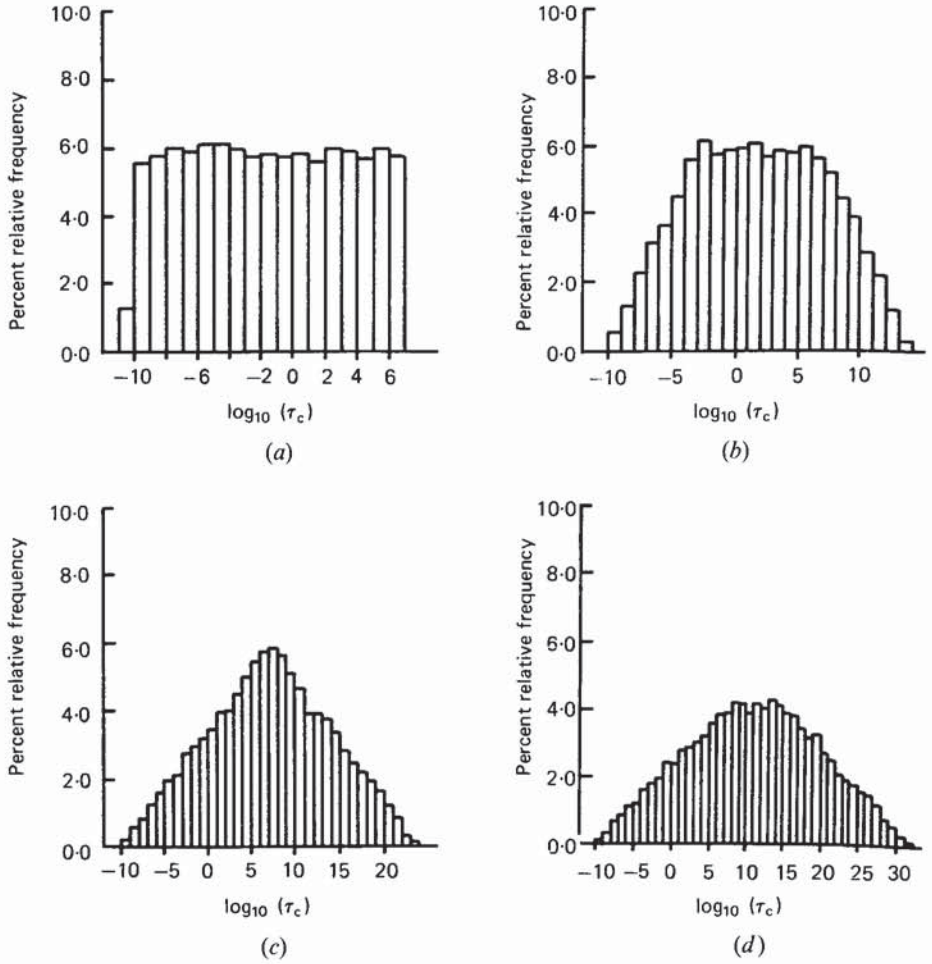


Figure 54. Distributions of trapping time constants evaluated using equation (7.7). The device parameters are appropriate to the device whose power spectrum is depicted in figure 21 (*a*): $V_D = 100\,\mathrm{mV}$, $I_D \approx 3{\cdot}5\,\mu\mathrm{A}$, $T = 293\,\mathrm{K}$. The range of ΔE_B is 0–1 eV. (*a*) $\sigma_0 = 10^{-15}\,\mathrm{cm}^2$; (*b*) $\sigma_0 = 10^{-22}$–$10^{-15}\,\mathrm{cm}^2$; (*c*) $\sigma_0 = 10^{-32}$–$10^{-15}\,\mathrm{cm}^2$; (*d*) $\sigma_0 = 10^{-40}$–$10^{-15}\,\mathrm{cm}^2$.

maximum cross-section is therefore $10^{-15}\,\mathrm{cm}^2$, corresponding, more or less, to a defect situated directly at the interface. Assuming a characteristic length, $x_0 \approx 0{\cdot}1\,\mathrm{nm}$ and a uniform distribution of defects up to about 5 nm into the oxide, we then find that a reasonable estimate for a lower bound on σ_0 is $\exp(-50) \times 10^{-15} \approx 10^{-40}\,\mathrm{cm}^2$. From figure 54 (*d*) it is apparent that the lower limit on σ_0 is of no real physical significance, since it is only time constants up to about a day or so that matter. Using equation (7.3) and the set of time constants shown in figure 54 (*d*), we obtain the power spectrum shown in figure 55 (*a*). In producing this figure, we have kept to our earlier observation of roughly one RTS per decade in time in a small-area device. An important point to note from figure 55 (*a*) is that the magnitude of the noise power spectral density is in reasonable agreement with the experimentally determined spectrum of figure 21 (*a*).

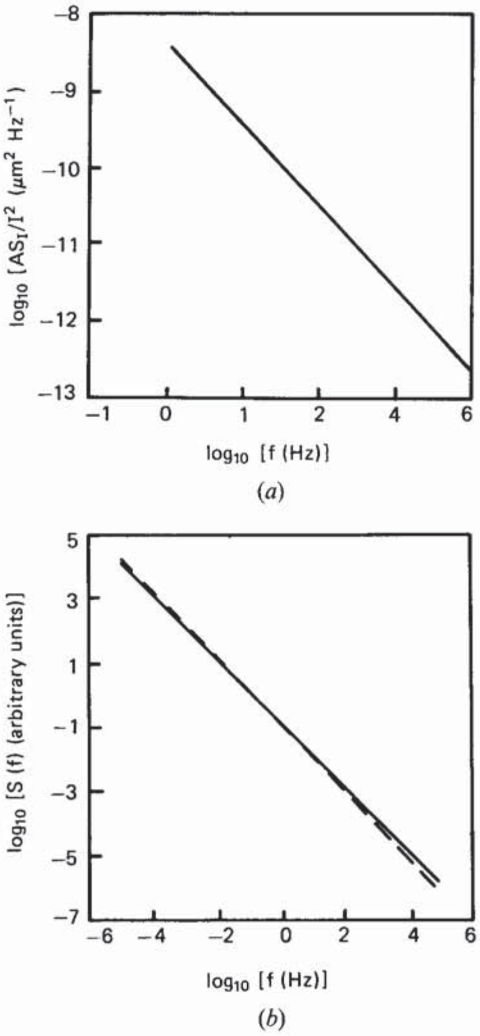


Figure 55. (*a*) Simulated power spectrum of a large-area MOSFET obtained using the distribution of time constants shown in figure 54 (*d*) and equation (7.3). This should be compared with figure 21 (*a*). (*b*) Power spectra generated using equations (7.10) and (7.11). $t_1 = 10^{-8}$ s and $t_2 = 10^{8}$ s. k (the intercept on the log τ axis of figure 54 (*d*) takes on the two values: $-\infty$ (full line) and -10 (dashed line). The former value gives a slope of $-1{\cdot}0$ and the latter $-1{\cdot}07$. Thus a distribution of time constants deviating substantially from the standard uniform distribution in log time gives a respectable $1/f$ spectrum.

We have thus arrived at the somewhat surprising result that a distribution of time constants (figure 54 (*d*)) deviating substantially from the standard uniform distribution in log time gives a respectable $1/f$ spectrum. To see how this comes above, consider the following argument. With respect to figure 54 (*d*), draw a line through the top of the histogram bins between log times -10 and 5, say. Let k be the intercept of this line on the log τ axis and m the intercept on a vertical line drawn through the point log $\tau = 0$. Then the total number of RTSs (traps) with time constants between t_1 and

t_2 ($-10 < \log t_1 < \log t_2 < 5$) is given by (area of trapezium)

$$N(t_1, t_2) = \frac{1}{2}\left[\left(-\frac{m}{k}\log t_1 + m\right) + \left(-\frac{m}{k}\log t_2 + m\right)\right] \times (\log t_2 - \log t_1). \qquad (7.9)$$

This may be written as

$$N(t_1, t_2) = \left[m \log \tau - \frac{m}{2k}(\log \tau)^2\right]_{t_1}^{t_2}.$$

Now we know that

$$N(t_1, t_2) = \int_{t_1}^{t_2} n(\tau)\, \mathrm{d}\tau,$$

and we can therefore write

$$n(\tau) = \frac{m}{2{\cdot}303\tau} - \frac{m}{2{\cdot}203^2 k\tau} \ln \tau. \qquad (7.10)$$

As $k \to -\infty$, the second term on the right-hand side of equation (7.10) vanishes, and we recover the result corresponding to a uniform distribution of time constants in log time, as required. It is the value of k in relation to the log of the time interval being integrated over that gives rise to any deviations from $1/f$ behaviour. Converting the sum in equation (7.1) to an integral, using equations (7.5) and (2.19 *b*), and assuming a mean amplitude ΔI, we find

$$S(f) \approx \frac{1}{N}\int_{t_1}^{t_2} \frac{n(\tau) 2(\Delta I)^2 \tau \,\mathrm{d}\tau}{4 + (2\pi f \tau)^2}. \qquad (7.11)$$

Using equation (7.11) in conjunction with equation (7.10), we have evaluated the power spectrum arising from a value of k compatible with the intercept in figure 54 (*d*) and a value of $k = -\infty$ (corresponding to a uniform distribution): see figure 55 (*b*). It is clear that a value of $k = -\infty$ gives a slope of -1, while a value of $k = -10$ still gives a respectable $1/f$ spectrum with a slope of around $-1{\cdot}07$. Thus a distribution of time constants of the form shown in figure 54 (*d*) gives a $1/f$ spectrum with a slope slightly greater than -1. However, in $1/f$ noise measurements slopes both less than and greater than -1 are observed. We have also investigated the effects of reducing still further the constraints on the distribution of activation energies that we have used hitherto. In particular, we have used Gaussian distributions and found the spectra to conform to a '$1/f$' form and to be in good agreement with experimental data.

To summarize, we have shown that the multiphonon model of carrier trapping into states in the oxide close to the Si/SiO_2 interface is consistent with the measured noise power spectral density of the current fluctuations in large-area MOSFETs. Although the present data set on individual defect states is still rather limited, we have shown that the form of the power spectrum is largely insensitive to the details of distributions of cross-section pre-factors and activation energies, provided that they are chosen in a physically meaningful way; though further work on the nature of the distributions is clearly called for. Moreover, one can conclude that measurement of the slope of a power spectrum alone is a very insensitive tool for determining the underlying physical mechanisms responsible for its generation. Finally, it is to be noted that defects in the bulk silicon would have well defined energy levels and cross-sections and would give rise to Lorentzian power spectra. Since we observe a $1/f$ spectrum, it is clear that the majority of the defects generating RTSs must reside in the oxide. As we saw in section 3.2.3, if the standard two-level RTSs are interpreted as multi-electron capture, this results in the traps having to be located in the silicon. Hence the observation of a $1/f$ spectrum represents indirect evidence that the majority of RTSs observed in small devices are caused by single-electron trapping.

7.2. *Conductance measurements on the Si/SiO$_2$ interface*

The four types of charge associated with the thermally grown oxide/silicon interface are the fixed oxide charge Q_f, the mobile ionic charge Q_m, the interface-state charge Q_{it} and the oxide trapped charge Q_{ot}. This is the approved standard nomenclature used to describe the various charges (Deal 1980). Mobile ionic charge is typically due to alkali-metal (e.g. Na) contamination of the oxide. Oxide trapped charge is usually located at the metal/SiO$_2$ interface or near the Si/SiO$_2$ interface. It is commonly produced by hot-electron injection or exposure to ionizing radiation. Q_f is a positive stable charge residing close to the Si/SiO$_2$ interface. It cannot be charged or discharged by varying the silicon surface potential, and its magnitude depends on the final oxidation or annealing. We shall not be concerned further with these types of charge; it is the interface-state charge that interests us. In particular, we wish to investigate the relationship between the interface states Q_{it} and the defect states responsible for $1/f$ noise and random telegraph signals.

The conventional definition for Q_{it} is as follows: Q_{it} is located at the interface and can be charged or discharged as the surface potential changes. The Q_{it} density (per unit area per electron volt) D_{it} can be minimized by a hydrogen anneal above 300°C. After the anneal, the residual D_{it} is constant near mid-gap, rising steadily towards the band edges. A typical mid-gap interface-state density in a modern process is about $10^{10}\,\mathrm{cm}^{-2}\,\mathrm{eV}^{-1}$, corresponding roughly to one interface state per 10^5 interface atoms. Thus the interface states are sufficiently well separated that they do not interact.

The main experimental methods used to investigate Q_{it} are the capacitance–voltage (C–V) and conductance (G–ω) techniques. We shall give a very brief outline of the G–ω method; full accounts of both methods are given by Nicollian and Brews (1982).

The conductance method has been fundamental to the present understanding of interface traps. It extracts interface trap-level density, capture probability and time-constant dispersion from the real component of the admittance of an MOS capacitor. Interface trap levels are detected through the loss resulting from changes in their occupancy produced by small variations of gate voltage. The conductance technique is simplest to understand in depletion because minority-carrier effects are not important; in depletion, interface traps change occupancy by capture and emission of majority carriers. A small a.c. voltage applied to the gate of an MOS capacitor moves the band edge at the interface relative to the Fermi level. Majority carriers are captured or emitted, changing the occupancy of the interface trap levels in a small energy interval a few kT wide centred about the Fermi level. This capture and emission of majority carriers causes an energy loss at all frequencies except the very lowest (to which all interface states immediately respond) and the very highest (to which no interface trap response occurs). Further discussion of the energy loss mechanism can be found in the work of Nicollian and Brews (1982), p. 180.