ADVANCES IN PHYSICS, 1989, VOL. 38, NO. 4, 367–468

Noise in solid-state microstructures: A new perspective on individual defects, interface states and low-frequency ($1/f$) noise

By M. J. KIRTON and M. J. UREN

Royal Signals and Radar Establishment, St Andrew's Road, Great Malvern,
Worcestershire WR14 3PS, U.K.

[Received 3 November 1988]

Abstract

In very small electronic devices the alternate capture and emission of carriers at an individual defect site generates discrete switching in the device resistance —referred to as a random telegraph signal (RTS). The study of RTSs has provided a powerful means of investigating the capture and emission kinetics of single defects, has demonstrated the defect origins of low-frequency ($1/f$) noise in these devices, and has provided new insight into the nature of defects at the Si/SiO_2 interface.

Contents

1. Introduction

The earliest observations of discrete-switching behaviour in the current flowing through semiconductor devices date back to the 1950s and 1960s (Buckingham 1983, Chapter 7); the phenomenon was referred to as burst noise and was usually associated with reverse-biased p–n junctions and bipolar transistors. However, the first clear demonstration that single-electron switching events were observable in electronic devices came from the pioneering study of Kandiah and Whiting (1978). They studied low-frequency noise in four-terminal silicon junction field-effect transistors (JFETs) and showed that it was dominated by Shockley–Read–Hall centres lying in the Debye region, between the channel and the fully depleted region. In good JFETs the number of such centres for devices of length 2 μm and width 1000 μm was between three and ten. The fluctuation of one unit charge at a centre electrostatically modulated the channel width; the fluctuations in channel width manifested themselves as measurable discrete changes in the source–drain current.

As a consequence of recent advances in processing technology, it has now become possible to produce devices in which the active volume is so small that it contains only a small number of charge carriers. The examples to which most of this article is addressed are small-area (about 1 μm^2) silicon metal–oxide–semiconductor field-effect transistors (MOSFETs) and metal–insulator–metal (MIM) tunnel junctions. As for the JFETs, in these devices the capture of an electron (or hole) into a localized defect state gives rise to a measurable change in device resistance. Figure 1 shows an example of the random telegraph signal (RTS) measured in the drain current of a MOSFET as a function of time; the times in the high- and low-current states correspond to carrier capture and emission respectively.

The bias-voltage dependence of the capture and emission times allows one to determine the location of the defects. In MOSFETs they are found (with a few provisos) to reside in the oxide up to a few nanometres from the interface, and hence within tunnelling distance of the inversion layer. For the MIM tunnel junctions, the traps are also located in the insulator. Through the study of the temperature and bias-voltage dependence of these times for a single defect, one can extract parameters such as capture cross-section, activation energy for capture and emission, and the temperature dependence of the trap energy level (trap entropy). These insulator defects are not found to be of a simple Shockley–Read–Hall type, but show evidence of strong lattice relaxation on capture as well as a large entropy change. In addition, they show a wide variation in all their characteristics such as energy level, capture activation energy and cross-section—quite consistent with their amorphous environment.

Another principal theme of this article is the relationship between these defects and the $1/f$ noise found in large devices. During the past two decades the origin of $1/f$ noise has been the subject of extensive investigation (Press 1978, Van der Ziel 1979, Dutta and Horn 1981, Hooge *et al.* 1981, Weissman 1988). Despite this intensive effort, the subject of $1/f$ noise has been notorious for several reasons: first, there has been a lack of data open to unambiguous interpretation; secondly, there has been a long-running and rather sterile debate over 'mobility-fluctuation' *versus* 'number-fluctuation' models; and thirdly, the quantum-$1/f$ theory of Handel and co-workers (Kousik *et al.* 1985) has aroused vitriolic debate between its opponents and supporters. The basic reason why no consensus has emerged is that little detailed information comes from the conventional ensemble-averaged power spectrum. We shall discuss the recent results on the noise properties of microstructures in which the averaging process is incomplete and individual fluctuators can be resolved. In the case of MOSFETs and MIM diodes, it will be shown conclusively that the $1/f$ noise in large devices is caused by the summation of many RTSs due to the defects in the insulator. In addition, the distribution of physical characteristics measured for the defects accounts easily for the wide range of time constants necessary to generate $1/f$ noise.

Whilst the majority of RTSs that are seen are of a form similar to that shown in figure 1, some signals show very complex switching behaviour including more than two levels or more than two time constants. Various explanations have been proposed, including collective capture into a defect cluster, Coulombic interaction within a defect cluster and physical reconfiguration within a set of metastable minima. We shall give various examples and discuss the class of models that can and cannot account for each one.

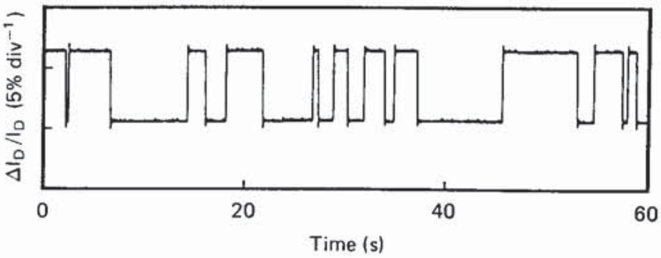


Figure 1. Random telegraph signal. Change in current against time. Active area of MOSFET is 0·4 μm^2, V_D = 10 mV, V_G = 0·94 V, I_D = 6·4 nA, T = 293 K. From Kirton *et al.* (1987).

The electrical activity of defects at the Si/SiO$_2$ interface is normally studied using capacitance–voltage or conductance–voltage techniques. The accepted model to account for the results has defect states distributed throughout the band gap, but with a single (smeared) time constant at any given energy. This is incompatible with the measured wide range of characteristics of the RTSs. Recent experiments that have used the conductance technique show that there are two classes of interface defect: the first includes those defects normally seen, and which presumably reside at the interface, and are characterized by a single time constant; the second class incorporates defects residing in the oxide, which have a wide range of time constants and are responsible for the RTSs and $1/f$ noise.

The remainder of this article is arranged as follows. In the next section we discuss those aspects of noise theory and defect theory that are essential for an understanding of this work. A short subsection is included on the practical measurement of noise spectra and RTSs. In section 3 we provide an overview of RTSs that have been observed in a variety of electronic devices, concentrating on the case of the silicon MOSFET. The decomposition of $1/f$ noise in microstructures into its constituent Lorentzian components is described in section 4. In section 5 the capture and emission kinetics of individual defects at the Si/SiO$_2$ interface and in the insulator of MIM tunnel junctions are discussed. Complex RTSs are considered along with defect metastability, non-Gaussian noise processes and collective carrier capture in section 6. In section 7 we consider the nature of trapping centres at the Si/SiO$_2$ interface and the way in which they give rise to a wide distribution of time constants and hence $1/f$ noise. In addition, we discuss the relationship between the $1/f$ states and the interface states commonly observed in capacitance–voltage and conductance measurements. We end with concluding comments in section 8.

2. Theoretical and experimental background

The main purpose of the following subsections is to provide a framework that will allow the reader to follow our detailed analysis of current noise, random telegraph signals (RTSs) and the capture and emission kinetics of individual defects. In this respect, we begin with a very brief account of the standard mathematical means of describing a purely random (stochastic) waveform, namely the autocorrelation function and the Wiener–Khintchine theorem (MacDonald 1962, Buckingham 1983). In sections 2.2 and 2.3 the probability distribution of an RTS and the form of its power spectrum are derived. Section 2.4 gives an outline of the practical measurement of RTSs and power spectra. We then end with a brief discussion in section 2.5 of the statistical mechanics of defect occupation.

2.1. *The autocorrelation function and the Wiener–Khintchine theorem*

The instantaneous values of a stochastic process cannot be predicted; thus the process must be characterized by its average statistical properties. One convenient statistical measure is the autocorrelation function, which records the memory or noisiness of the fluctuations. In a statistically stationary process (i.e. one in which the statistical properties do not depend on the epoch in which they are measured) the autocorrelation function $c(\tau)$ is defined by

$$c(\tau) \equiv \langle x(t)x(t+\tau)\rangle = \lim_{T\to\infty} \frac{1}{T}\int_{-T/2}^{T/2} x(t)x(t+\tau)\,\mathrm{d}t. \qquad (2.1)$$

The angular brackets denote a time average and T is the duration of the observation interval. Alternatively, the averaging can be carried out by summing over all members of an ensemble at a given time t. In either case, for a statistically stationary process, $c(\tau)$ is independent of t.

A strictly periodic function can be expanded in a Fourier series. In the limit in which the period tends to infinity the series expansion is replaced by an integral. One can use the Fourier integral formalism followed by ensemble averaging to determine the average spectral content of random fluctuations. We define the Fourier transform and its inverse by the following standard relationships:

$$X(\omega) = \int_{-\infty}^{\infty} x(t) \exp(-\mathrm{i}\omega t)\, \mathrm{d}t, \tag{2.2.a}$$

$$x(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} X(\omega) \exp(\mathrm{i}\omega t)\, \mathrm{d}\omega. \tag{2.2.b}$$

We assume that the fluctuations $x(t)$ are observed in the time interval $[-\frac{1}{2}T, \frac{1}{2}T]$ and they are zero outside this interval. Using Parseval's theorem

$$\int_{-\infty}^{\infty} x_1(t)x_2^*(t)\, \mathrm{d}t = \frac{1}{2\pi}\int_{-\infty}^{\infty} X_1(\omega)X_2^*(\omega)\, \mathrm{d}\omega,$$

we can write

$$\int_{-\infty}^{\infty} [x_T(t)]^2\, \mathrm{d}t = \frac{1}{2\pi}\int_{-\infty}^{\infty} |X_T(\omega)|^2\, \mathrm{d}\omega, \tag{2.3}$$

where the subscript T refers to the time interval in which the fluctuations are non-zero. Conventionally, the left-hand side of equation (2.3) is equal to the total energy of the fluctuations. The average power can be obtained by dividing equation (2.3) by T:

$$\lim_{T\to\infty} \frac{1}{T}\int_{-\infty}^{\infty} [x_T(t)]^2\, \mathrm{d}t = \lim_{T\to\infty} \frac{1}{2\pi}\int_{0}^{\infty} \frac{2|X_T(\omega)|^2\, \mathrm{d}\omega}{T}. \tag{2.4}$$

Since $x_T(t)$ is real, the integral on the right-hand side of equation (2.4) has been limited to positive frequencies and a factor of two incorporated. The term $2|X_T(\omega)|^2/T$ has dimensions of power per Hertz. Carrying out an ensemble average for this stochastic process, we obtain the average power spectral density:

$$S(\omega) = \lim_{T\to\infty} \frac{\overline{2|X(\omega)|^2}}{T}. \tag{2.5}$$

The practical measurement of $S(\omega)$ is outlined in section 2.4. Finally, the power spectral density is related to the autocorrelation function via the Wiener–Khintchine theorem (Buckingham 1983):

$$c(\tau) = \frac{1}{2\pi}\int_{0}^{\infty} S(\omega) \cos \omega\tau\, \mathrm{d}\omega, \tag{2.6 a}$$

$$S(\omega) = 4\int_{0}^{\infty} c(\tau) \cos \omega\tau\, \mathrm{d}\tau. \tag{2.6 b}$$

2.2. *Probability distribution of an RTS*

Referring back to figure 1, we shall take the high-current state of the RTS to be state 1 and the low-current state to be state 0. We shall assume that the probability (per unit time) of a transition from state 1 to state 0 (i.e. from up to down) is given by $1/\bar{\tau}_1$, with $1/\bar{\tau}_0$ being the corresponding probability from 0 to 1 (i.e. from down to up). The transitions are instantaneous. We now intend to show that these assumptions imply that the times in states 0 and 1 are exponentially distributed, that is, the switching is a Poisson process.

Let $p_1(t)\,\mathrm{d}t$ be the probability that state 1 will not make a transition for time t, then will make one between times t and $t + \mathrm{d}t$. Thus

$$p_1(t) = A(t)/\bar{\tau}_1, \tag{2.7}$$

where $A(t)$ is the probability that after time t state 1 will not have made a transition and $1/\bar{\tau}_1$ is the probability (per unit time) of making a transition to state 0 at time t. However,

$$A(t + \mathrm{d}t) = A(t)(1 - \mathrm{d}t/\bar{\tau}_1); \tag{2.8}$$

that is, the probability of not making a transition at time $t + \mathrm{d}t$ is equal to the product of the probability of not having made a transition at time t and the probability of not making a transition during the interval from t to $t + \mathrm{d}t$. We can rearrange equation (2.8) to give

$$\frac{\mathrm{d}A(t)}{\mathrm{d}t} = -\frac{A(t)}{\bar{\tau}_1}. \tag{2.9}$$

Integrating both sides of equation (2.9), we find

$$A(t) = \exp(-t/\bar{\tau}_1), \tag{2.10}$$

such that $A(0) = 1$. Thus

$$p_1(t) = \frac{1}{\bar{\tau}_1}\exp\left(-\frac{t}{\bar{\tau}_1}\right). \tag{2.11 a}$$

$p_1(t)$ is correctly normalized such that

$$\int_0^\infty p_1(t)\,\mathrm{d}t = 1.$$

The corresponding expression for $p_0(t)$ is

$$p_0(t) = \frac{1}{\bar{\tau}_0}\exp\left(-\frac{t}{\bar{\tau}_0}\right). \tag{2.11 b}$$

Hence, on the assumption that the up and down times are characterized by single attempt rates, we expect the times to be exponentially distributed. The mean time spent in state 1 is given by

$$\int_0^\infty t p_1(t)\,\mathrm{d}t = \bar{\tau}_1, \tag{2.12 a}$$

and the standard deviation is

$$\left[\int_0^\infty t^2 p_1(t)\,\mathrm{d}t - \bar{\tau}_1^2\right]^{1/2} = \bar{\tau}_1. \tag{2.12 b}$$

Equivalent expressions hold for the down state. Thus the standard deviation is equal to the mean time spent in the state. Equation (2.12 *b*) can be used as a simple test for exponential behaviour. We return to this point again in section 3.2.

2.3. *Power spectrum of an RTS: Lorentzian spectrum*

Here we shall outline Machlup's (1954) derivation of the power spectrum of an asymmetric RTS. Initially, we need to evaluate the autocorrelation function of the RTS. It is convenient to choose the origin of the coordinate system such that state 0 has amplitude $x_0 = 0$, and state 1 has amplitude $x_1 = \Delta I$. In addition, all statistical properties will be taken to be independent of the time origin. The probability that at any given time the RTS is in state 1 is $\bar{\tau}_1/(\bar{\tau}_0 + \bar{\tau}_1)$, and similarly for state 0 it is $\bar{\tau}_0/(\bar{\tau}_0 + \bar{\tau}_1)$. Then we have

$$\begin{aligned} c(t) &= \langle x(s)x(s+t)\rangle \\ &= \sum_i \sum_j x_i x_j \times \{\text{Prob. that } x(s) = x_i\} \\ &\qquad \times \{\text{Prob. that } x(s+t) = x_j, \text{ given } x(s) = x_i\}. \end{aligned} \tag{2.13}$$

Since $x_0 = 0$ and $x_1 = \Delta I$, we obtain

$$\begin{aligned} c(t) &= (\Delta I)^2 \frac{\bar{\tau}_1}{\bar{\tau}_0 + \bar{\tau}_1} P_{11}(t) \\ &= (\Delta I)^2 \times \{\text{Prob. that } x(s) = \Delta I\} \\ &\qquad \times \{\text{Prob. of even no. of transitions in time } t, \text{ starting in state } 1\}. \end{aligned} \tag{2.14}$$

If we define $P_{10}(t)$ as the probability of an odd number of transitions in time t, starting in state 1 then we have

$$P_{11}(t) + P_{10}(t) = 1. \tag{2.15}$$

In addition,

$$P_{11}(t + \mathrm{d}t) = P_{10}(t)\frac{\mathrm{d}t}{\bar{\tau}_0} + P_{11}(t)\left(1 - \frac{\mathrm{d}t}{\bar{\tau}_1}\right); \tag{2.16}$$

that is, the probability of an even number of transitions in time $t + \mathrm{d}t$ is given by the sum of two mutually exclusive events: first, the probability of an odd number of transitions in time t and one transition in time $\mathrm{d}t$; and secondly, the probability of an even number of transitions in time t and no transitions in time $\mathrm{d}t$. We can make $\mathrm{d}t$ small enough that the probability of more than one transition is vanishingly small. Substituting from equation (2.15) into equation (2.16), we obtain the following differential equation for $P_{11}(t)$:

$$\frac{\mathrm{d}P_{11}(t)}{\mathrm{d}t} + P_{11}(t)\left(\frac{1}{\bar{\tau}_0} + \frac{1}{\bar{\tau}_1}\right) = \frac{1}{\bar{\tau}_0}. \tag{2.17}$$

This equation can be solved by using $\exp\left[\int (1/\bar{\tau}_0 + 1/\bar{\tau}_1)\,\mathrm{d}t\right]$ as an integrating factor:

$$P_{11}(t) = \frac{\bar{\tau}_1}{\bar{\tau}_0 + \bar{\tau}_1} + \frac{\bar{\tau}_0}{\bar{\tau}_0 + \bar{\tau}_1}\exp\left[-\left(\frac{1}{\bar{\tau}_0} + \frac{1}{\bar{\tau}_1}\right)t\right], \tag{2.18}$$

where $P_{11}(t)$ has been normalized such that $P_{11}(0) = 1$. Equations (2.18), (2.14) and (2.6 *b*) can now be used to evaluate the power spectral density $S(f)$:

$$S(f) = \frac{4(\Delta I)^2}{(\bar{\tau}_0 + \bar{\tau}_1)[(1/\bar{\tau}_0 + 1/\bar{\tau}_1)^2 + (2\pi f)^2]}. \tag{2.19 a}$$

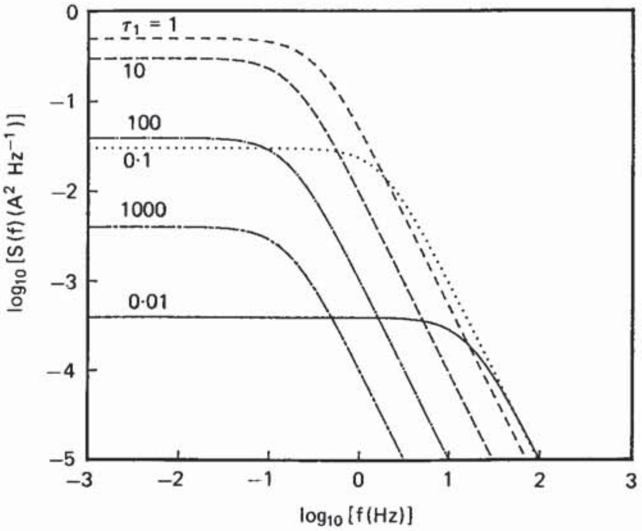


Figure 2. Lorentzian power spectra calculated using equation (2.19 *a*). $\Delta I = 1$ A, $\tau_0 = 1{\cdot}0$ s, $\tau_1 = 0{\cdot}01$–1000 s.

(The d.c. term, which contributes a delta function at $f = 0$, has been ignored.) For the case of a symmetric RTS, that is, $\bar{\tau}_0 = \bar{\tau}_1 = \bar{\tau}$ for example, this equation simplifies to

$$S(f) = \frac{2(\Delta I)^2 \bar{\tau}}{4 + (2\pi f \bar{\tau})^2}. \tag{2.19 b}$$

The total power P in the RTS can be obtained by integrating equation (2.19 *a*) over all frequencies:

$$P = \frac{(\Delta I)^2}{(\bar{\tau}_0 + \bar{\tau}_1)(1/\bar{\tau}_0 + 1/\bar{\tau}_1)}. \tag{2.20}$$

As one would expect, $P = (\frac{1}{2}\Delta I)^2$ when $\bar{\tau}_0 = \bar{\tau}_1$; P is a maximum under these conditions.

In figure 2 we have plotted $S(f)$ against frequency for the cases in which $\Delta I = 1{\cdot}0$ A, $\bar{\tau}_0 = 1{\cdot}0$ s and $\bar{\tau}_1$ takes on six values between 0·01 s and 1000 s. These are all characteristic Lorentzian spectra. Starting at $\bar{\tau}_1 = 0{\cdot}01$ s, we find the turnover frequency around 10 Hz and a small value for the total integrated power. As $\bar{\tau}_1$ increases, the turnover frequency steadily decreases as the power increases. The condition of maximum power is reached at $\bar{\tau}_1 = \bar{\tau}_0 = 1{\cdot}0$ s. Thereafter, as $\bar{\tau}_1$ takes on the values 10 up to 1000 s, the power decreases and the turnover frequency tends to a constant.

2.4. *Practical measurement of RTSs and power spectra*

Telegraph signals are characterized by four parameters: the averages of the up and down times, the magnitude of the telegraph signal and the size of the background upon which the signal is superimposed. All of these will vary over wide ranges

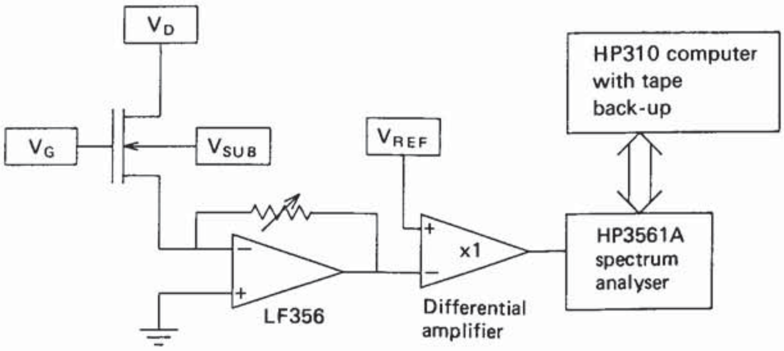


Figure 3. Block diagram of experimental set-up used for the measurement of power spectra and random telegraph signals in MOSFETs.

as the device temperature and bias voltages are changed, so a measurement system of great flexibility is necessary. There are obviously many possible approaches, each optimized for a particular sample type, but figure 3 shows the particular approach that we adopted for our measurements of small MOSFETs around room temperature.

The MOSFET is contained in a small computer-controlled cryostat (MMR system 1, which is based on Joule–Thompson cooling using high-pressure nitrogen). The MOSFET bias voltages (V_G, V_D and V_{SUB}) are all computer-controlled and low-pass filtered; the drain supply is put through an extra potential divider to give additional stability. The source current is measured with a virtual-earth amplifier (a selected LF356 op-amp) with computer-switched feedback resistors giving $10^{-8}\,\mathrm{A\,V^{-1}}$ to $10^{-5}\,\mathrm{A\,V^{-1}}$ sensitivity, with a bandwidth extending out to 10^5 Hz. Since the signal is going to be digitized, it is essential to preserve dynamic range, so the background must be subtracted. This is done with a × 1/× 10 differential amplifier where V_{REF} is adjusted (under computer control) to give a zero average input signal into the A/D converter in the spectrum analyser. A major advantage in using a spectrum analyser to acquire data is that it has built in anti-aliasing filters that track the sampling rate and limit the bandwidth. Hence an optimum signal-to-noise ratio is obtained at all times.

To obtain a reasonable estimate (10% error) of the up and down times of the RTS requires that a time record containing more than 200 transitions is stored. Similarly, in order not to miss transitions, the sampling rate must be at least 100 × the average up or down time, implying a minimum time-record length of 20 000 points. The particular spectrum analyser that we used for the data capture (HP3561A) has a time buffer 40 960 points long, which in practice was found to be barely adequate. Since to measure, say, an activation energy requires typically ten temperatures, large amounts of data are quickly generated and must be stored and analysed in a convenient fashion. This analysis process is time-consuming because it is unusual to find traces that contain only one active trap. The signal-to-noise ratio is rarely good enough to allow automatic processing, so an initial inspection of the data is carried out by eye and any occasional spikes or transitions due to other traps removed before a computer program determines the average up and down times.

The system shown in figure 3 is also capable of measuring the noise power spectral density of the source current. The FFT spectrum analyser averages typically 100 spectra to give a reasonable (10%) estimate of the noise power. The procedure used was to measure the noise with no drain bias, which represents the Nyquist noise of the sample and the system noise. The drain bias was then applied, thus generating excess noise, and the noise measured again. Subtraction of the system noise gave the excess-current noise of the sample. This procedure was obviously only suitable for small drain biases where the device is still linear.

2.5. *Occupancy levels and the grand partition function*

In order to be precise in our meaning, we shall introduce the nomenclature 'occupancy level' $E(n + 1/n)$ to describe the energy level of a defect: $E(n + 1/n)$ marks the Fermi level E_F at which the defect's occupancy changes from n electrons to $n + 1$ electrons. (In our discussion of two-electron capture in section 6 we shall find this notation very beneficial to the analysis of the problem.) We can determine the occupany of the defect using the grand partition function, Z_G. This is written as (Kittel and Kroemer 1980)

$$Z_G = \sum_{ASN} \exp\left(-\frac{E_S - NE_F}{kT}\right), \tag{2.21}$$

where A S N implies the summation is to be carried out over all states S of the system for all numbers of particles N. We have adopted the convention of semiconductor physics and set E_F to be equivalent to the temperature-dependent chemical potential. The absolute probability that the system will be found in a state (N_1, E_1) is given by

$$p(N_1, E_1) = \frac{\gamma \exp[-(E_1 - N_1E_F)/kT]}{Z_G}, \tag{2.22}$$

where the state is orbitally (and perhaps also spin) degenerate with degeneracy γ.

Consider a defect system that has only two states of charge, n and $n + 1$, available. Let the energy zero of the system correspond to the defect occupied by n electrons. Then

$$Z_G = \gamma(n) \exp\left(\frac{nE_F}{kT}\right) + \gamma(n + 1) \exp\left[-\frac{E(n + 1/n) - (n + 1)E_F}{kT}\right], \tag{2.23}$$

where $\gamma(n)$ and $\gamma(n + 1)$ are the degeneracies of the n- and $(n + 1)$-electron states. Then the probability of finding the defect in the $(n + 1)$-electron state is

$$f = p(n + 1) = \left\{1 + g \exp\left[\frac{E(n + 1/n) - E_F}{kT}\right]\right\}^{-1}, \tag{2.24}$$

where

$$g = \gamma(n)/\gamma(n + 1). \tag{2.25}$$

This looks like a Fermi–Dirac distribution with a degeneracy factor g. In addition, we can write

$$\frac{p(n + 1)}{p(n)} = \frac{\gamma(n + 1)}{\gamma(n)} \exp\left[-\frac{E(n + 1/n) - E_F}{kT}\right]. \tag{2.26}$$

That is, when the Fermi level crosses the level $E(n + 1/n)$ the $(n + 1)$-electron state dominates over the n-electron state.

For an individual RTS generated by a trap with occupancy level $E(n + 1/n)$ and with mean capture and emission times $\bar{\tau}_c$ and $\bar{\tau}_e$, we have

$$f = \frac{\bar{\tau}_e}{\bar{\tau}_c + \bar{\tau}_e} = \left\{1 + g \exp\left[\frac{E(n + 1/n) - E_F}{kT}\right]\right\}^{-1}, \qquad (2.27\,a)$$

$$\bar{\tau}_e = \frac{\bar{\tau}_c}{g} \exp\left[-\frac{E(n + 1/n) - E_F}{kT}\right], \qquad (2.27\,b)$$

where g is given by equation (2.25).

3. Random telegraph signals and single-electron switching events

We begin with an overview of the diverse range of electronic devices and systems in which discrete-switching behaviour in the resistance has been observed. The specific example of the small-area silicon MOSFET is then taken. We show that the experimental data is consistent with an RTS being generated through the capture and emission of a single electron at an individual defect residing in the oxide close to the Si/SiO_2 interface. The mechanisms whereby a trapped electron modulates the channel conductivity still remain uncertain.

3.1. *Overview of RTS phenomena*

The so-called burst noise in reverse-biased p–n junctions and bipolar transistors provides us with our first example of discrete switching behaviour in electronic devices. Figure 4 shows the temporal variation in current through a reverse-biased germanium p–n junction recorded by Wolf and Holler (1967). They found that the times in the high- and low-current states obeyed Poisson statistics. In addition, they noted that the switching rate was thermally activated, with an activation energy larger than the germanium band gap. Although first observed nearly thirty years ago, the origins of burst noise still remain uncertain; dislocations, metal precipitates and the switching on and off of surface conduction channels have all been implicated in its production. Buckingham (1983) gives an excellent review of the phenomenon.

Definitive studies of RTS behaviour were first carried out by Kandiah and coworkers (Kandiah and Whiting 1978, Kandiah *et al.* 1981). They observed switching behaviour in double-gated silicon JFETs; see figure 5. Through very careful experimentation, they were able to show that the RTSs were generated through the charging

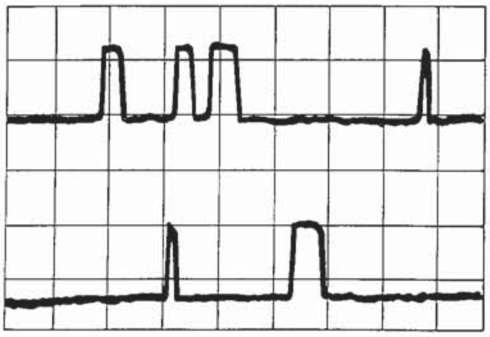


Figure 4. Waveform of burst noise in a reverse-biased (7·5 V) germanium junction. The horizontal scale is 20 ms div^{-1} and the vertical scale 20 nA div^{-1}. After Wolf and Holler (1967). © American Institute of Physics. Reproduced with permission.

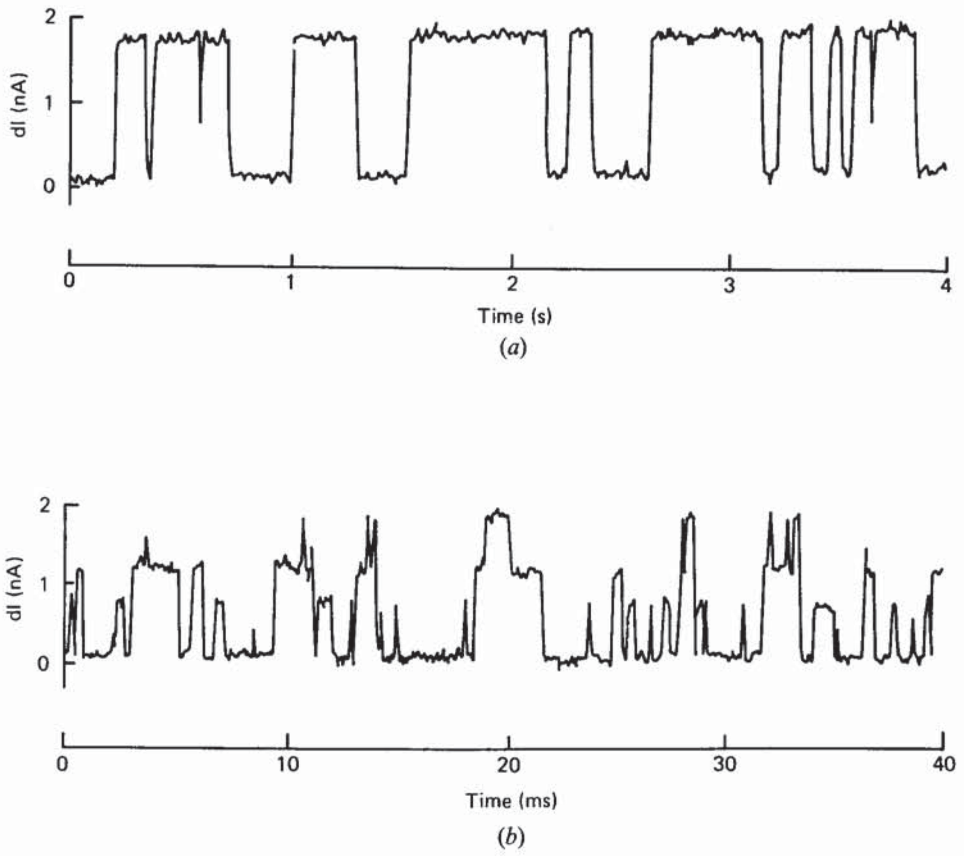


Figure 5. Current switching in silicon JFET, XS01: (*a*) $T = 83\,\text{K}$, $V_D = 6\,\text{V}$, $I_D = 16\,\mu\text{A}$; (*b*) $T = 166\,\text{K}$, $V_D = 2\,\text{V}$, $I_D = 17\,\mu\text{A}$. $V_G = +0{\cdot}44\,\text{V}$. Figure courtesy of K. Kandiah.

and discharging of single defects in the Debye region, between the channel and fully depleted region. In its more negative charge state and for n-channel devices, the defect produced a constriction of the channel; on electron emission, the channel resistance was lowered and the source–drain current increased. The extra gate allowed the channel to be moved in a direction normal to the current flow, and thus for defects to be moved into and out of the active region.

In a seminal paper, Ralls *et al.* (1984) studied fluctuation phenomena in very small MOSFETs (dimensions $0{\cdot}1\,\mu\text{m} \times 1{\cdot}0\,\mu\text{m}$) operating at cryogenic temperatures. They noted that as the device area was scaled down, the total number of Si/SiO_2 interface defects was correspondingly reduced. In small enough devices it is quite likely that only a handful of traps will have energy levels within kT or so of the surface Fermi level and thus will be fluctuating in occupancy. This is borne out in figure 6, which shows some of their results for several gate voltages and temperatures. The observed resistance changes are consistent with a single electron being removed from the channel and captured in a localized defect state. Note that as the gate voltage changes the mark-space ratio changes as the separation of the trap energy level and surface Fermi level is altered. It is also quite clear that the switching rate is a sensitive function

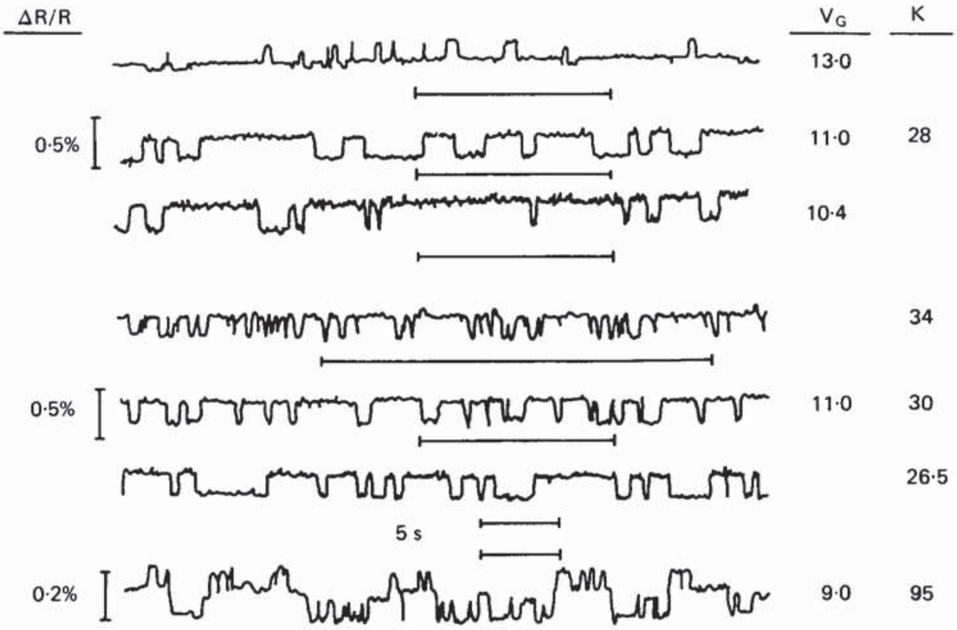


Figure 6. Resistance switching observed in a small ($0{\cdot}15\,\mu m \times 1{\cdot}0\,\mu m$) silicon MOSFET. From Ralls *et al.* (1984). © American Physical Society. Reproduced with permission.

of temperature. In addition, one can see that at elevated temperatures, where several RTSs are active, the resistance fluctuations are beginning to resemble the trace one would observe for a $1/f$ noise source.

Karwath and Schulz (1988) have applied techniques from deep-level transient spectroscopy (DLTS) to the study of small MOSFETs. First they applied a voltage pulse of several volts to the gate for a few minutes to fill the interface traps. They then pulsed the device into weak inversion and observed the emission of carriers from filled interface states and single-electron switching in the drain-voltage transient (figure 7). This technique allows the simultaneous measurement of all the traps in the accessible range of E_F, but is rather hard to analyse. Investigations of switching and interface-state generation induced in small MOSFETs by hot-carrier stress have been carried out by Bollu *et al.* (1987).

Judd *et al.* (1986) have observed switching phenomena in the tunnelling current through large-area ($180\,\mu m^2$) GaAs n–i–n diodes, where the intrinsic region consists of a linearly-graded band gap of AlGaAs: see figure 8. It is clearly worth noting that their results bear a strong resemblence to the aforementioned burst-noise work. They concluded that the switching was controlled by a single defect in the barrier region. Close to turn-on, the I–V characteristic will be dominated by inhomogeneities in the barrier. Any defect residing close to such an inhomogeneity will have a dramatic effect as it charges and discharges, leading to filamentary current transport and a measurable switching effect.

Our final example comes from the work of Welland and Koch (1986), who used the scanning tunnelling microscope (STM) (Binnig and Rohrer 1986) to study thin ($1{\cdot}5$–$2{\cdot}0$ nm) layers of thermally grown SiO_2. With the STM operating in the

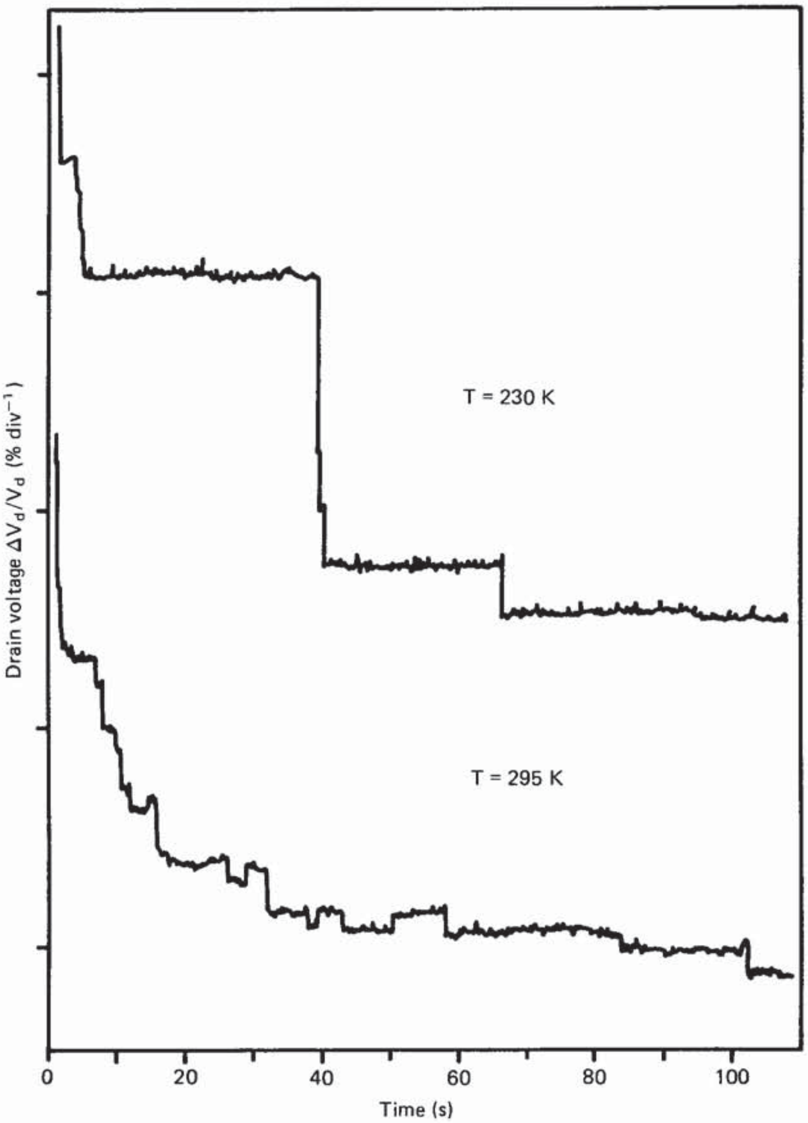


Figure 7. DLTS transients of electron re-emission in an n-channel MOSFET (1·2 μm × 1·5 μm) after complete filling of interface traps by a gate voltage of 4·4 V for 5 min. V_G = 0·85 V, I_D = 20 nA. From Karwath and Schulz (1988). © American Institute of Physics. Reproduced with permission.

constant-current mode, Welland and Koch found regions of the surface where current transients were faster than the response time of the feedback loop. These regions were further investigated and, with the feedback loop set to maintain an average current of 1 nA, the instantaneous current was found to switch between levels at 2 nA and close to zero. Figure 9 shows the tunnel current against time for various values of the potential on the tip of the probe and as the tip is moved away from the site of bistable-current production. Welland and Koch concluded that their data were

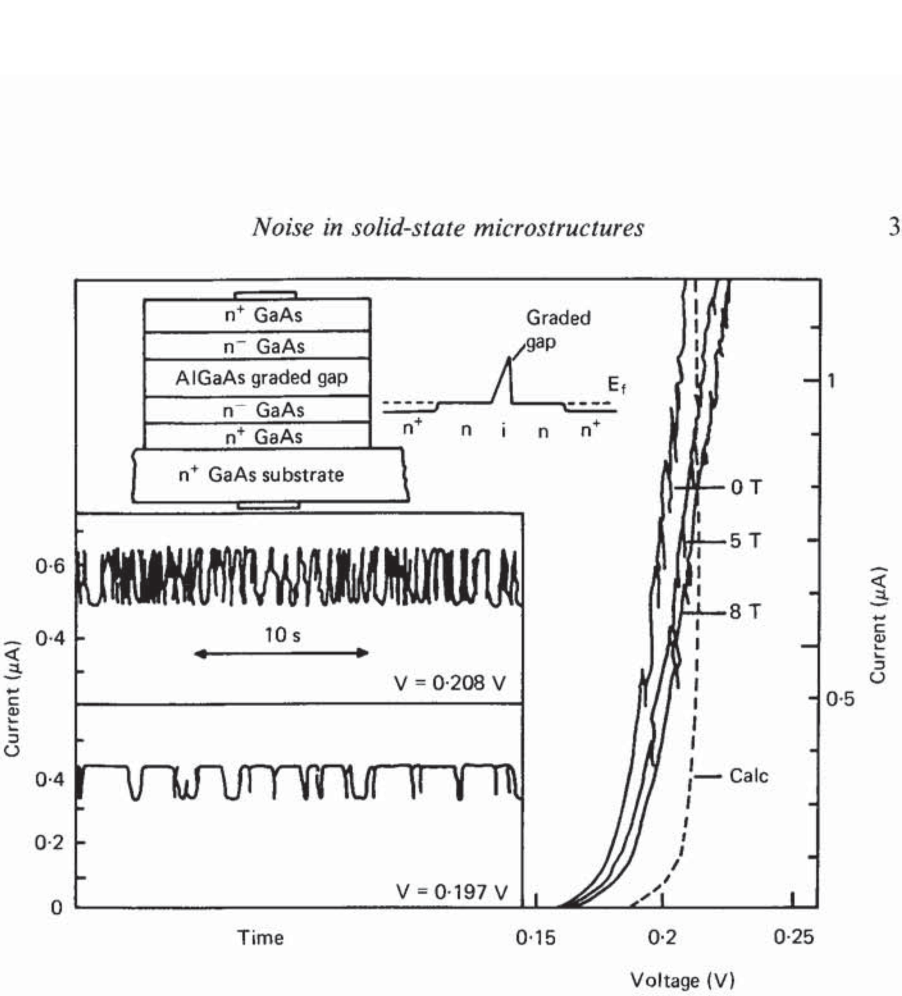


Figure 8. Current against forward bias voltage at 4·2 K. The solid lines are experimental traces at magnetic fields of 0, 5 and 8 T. The dashed line is the calculated response. The upper inset shows a schematic diagram of the structure. The lower inset shows the current as a function of time for two constant voltages. From Judd *et al.* (1986). © American Institute of Physics. Reproduced with permission.

consistent with electrons filling and leaving localized states on the surface. They also estimated the area surrounding the defect in which the tunnelling current was perturbed by the presence of the trapped charge. They found sufficiently close agreement between theory and experiment to conclude that the fluctuations were brought about by single-electron capture and emission events.

3.2. *Single-electron switching in small-area MOSFETs*

It is to be hoped that section 3.1 has given the reader a feel for the situations in which current switching occurs. Using the silicon MOSFET as an example, we shall now show that in this system single-electron trapping at interface defects provides a consistent explanation for the measured properties of the RTSs. We begin with a discussion of the device characteristics of the small MOSFETs used in our own investigations. This is followed by a consideration of the RTSs' gate-voltage dependence and the distribution of capture and emission times. In section 3.2.3 we consider in some detail the behaviour of the amplitudes of the RTSs both as a function of channel conductivity and temperature.

3.2.1. *Device characteristics*

The basic operation of MOSFETs is well understood (see Sze (1981)), and we shall only discuss some of the aspects pertinent to our measurements. The MOSFETs used

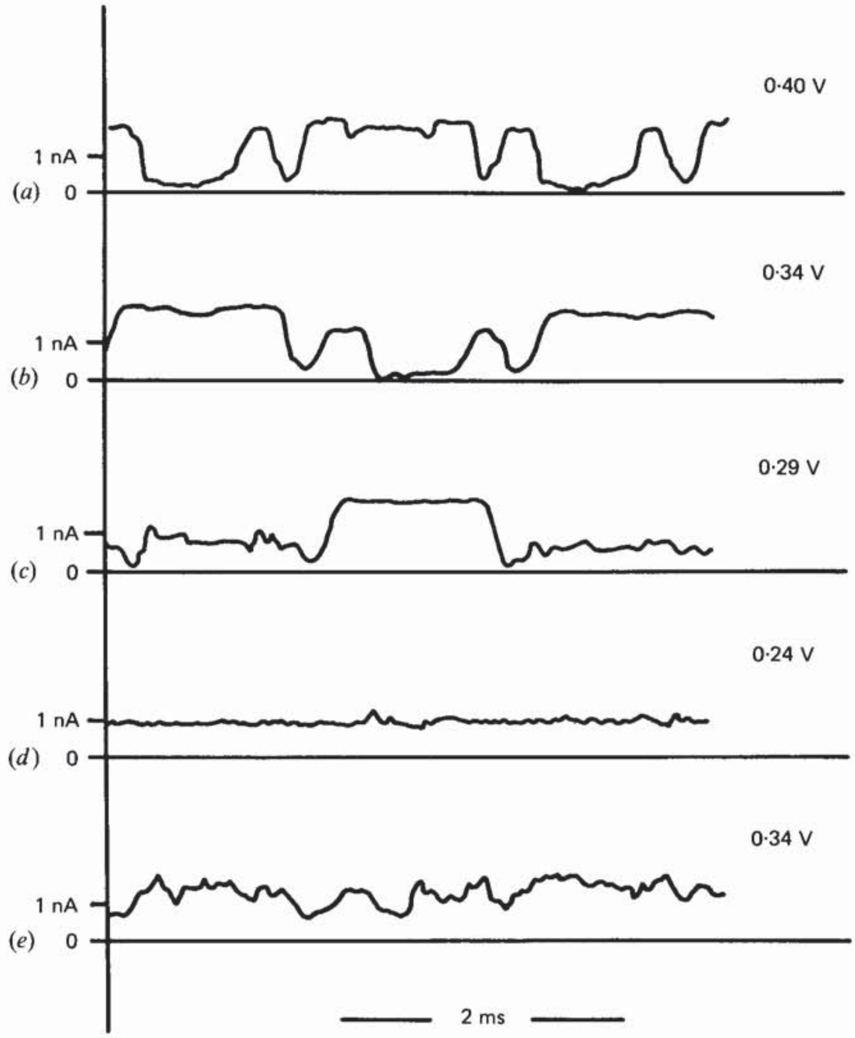


Figure 9. (*a*)–(*d*) STM tunnel current through thin SiO_2 layer as a function of time for four values of V_{tip}. The tip is positioned over peak of noise source. (*e*) The tunnel current against time for V_{tip} at 0·340 V and the tip positioned 3 nm away from the peak of the noise source. After Welland and Koch (1986). © American Institute of Physics. Reproduced with permission.

in all our studies came from processes that used a LOCOS (*LOC*al *O*xidation of *S*ilicon) technique to isolate the edges of the devices. This involves the definition of an active device area using a diffusion mask of silicon nitride; subsequent oxidation, removal of the nitride and then further oxidation produces a thick field oxide surrounding and isolating the thin device gate oxide. A schematic cross-section of the transistor across its width is shown in figure 10 (*c*), where it can be seen that the channel width is less than the dimension of the original mask owing to lateral encroachment of the thick field oxide. Similarly, figure 10 (*b*) shows a cross-section along the length of the device where the physical dimension of the gate is an

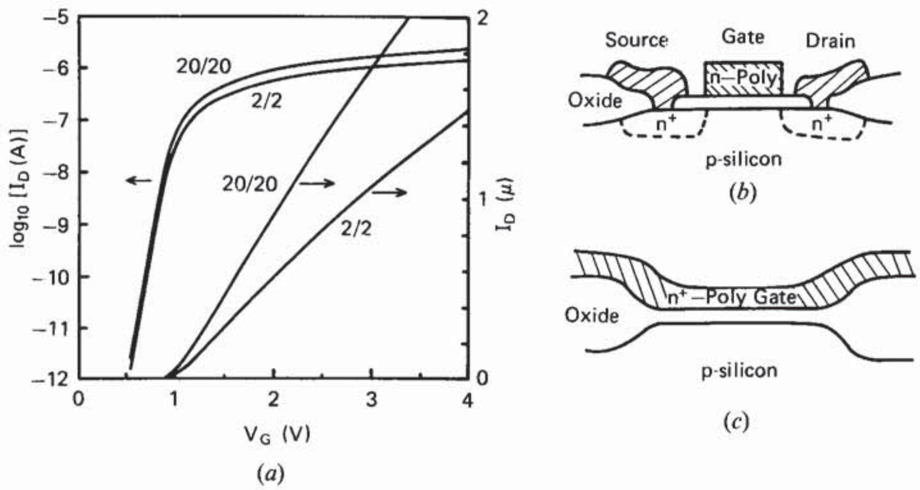


Figure 10. (*a*) Plots of I_D against V_G on both log and linear axes for 20 μm × 20 μm and 2 μm × 2 μm n-channel MOSFETs. V_D = 20 mV. The characteristics of the small device are described well by the long-channel model. (*b*) Cross-section of device structure along line joining source and drain. (*c*) Cross-section of device along line through middle of gate perpendicular to source–drain direction.

overestimate of the electrical channel length owing to sideways diffusion of the source and drain implants and the width of their depletion regions. In our samples we measured these corrections to the nominal dimensions by measuring the gain of various length and width transistors at low drain voltage and extrapolating to find each correction. Obviously, there will be some gate-voltage dependence to this true electrical width and length, but this was assumed to be small.

The majority of our measurements were carried out on devices from a single wafer manufactured at the GEC Hirst Research Centre, Wembley, U.K. The process was designed for a minimum geometry of 2·5 μm, but on the test mask there were devices of nominal 2 μm × 2 μm. After correcting the length and width, we found these devices to have effective electrical dimensions of 0·5 μm wide by 0·75 μm long. Small devices of this sort show so-called 'short-channel effects' due to the drain-voltage dependence of the drain depletion-region width. If operated at low drain voltage ($\leqslant$ 100 mV), these transistors show characteristics that are sufficiently similar to their long-channel counterparts that they can be modelled using simple long-channel models (Pao and Sah 1965, Brews 1978). Figure 10 (*a*) shows the gate-voltage characteristics for a small and a large transistor. The sub-threshold slopes, which are a sensitive test for deviations from long-channel behaviour, were found to be very similar: 87 mV decade^{-1} for the 20 μm × 20 μm device and 92 mV decade^{-1} for the 2 μm × 2 μm device. Hence, throughout our work, long-channel models were used to characterize the transistors.

3.2.2. *Gate-voltage dependence of RTSs*

The dependence on gate voltage of an RTS measured in a 0·4 μm^2 n-channel MOSFET at room temperature is shown in figure 11. This figure shows that as the gate voltage is increased, the time in the high-current state is reduced dramatically, while the time in the low-current state appears to be largely unaffected. Figure 12

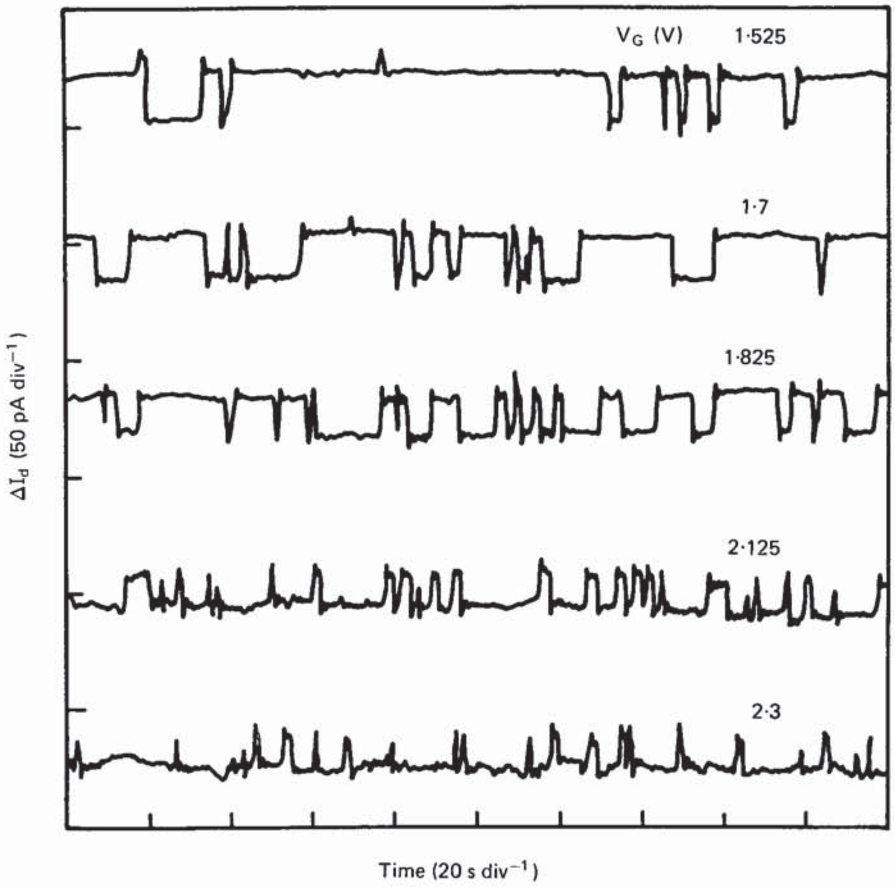


Figure 11. Random telegraph signals in small MOSFET measured at the indicated gate voltages. Active device area is 0·4 μm^2, V_D = 4 mV, T = 293 K. From Uren *et al.* (1985).

shows a schematic representation of the band bending in a (small-area) MOSFET in which there is only one defect energy level E_T within kT of the surface Fermi level E_F; the effect of a positive increment in gate voltage is shown by the dotted line. The important point to note is that the energy separation $E_T - E_F$ becomes less positive (or, with E_T below E_F, more negative) as V_G increases. For the linear regime of MOSFET operation the fractional occupancy of the defect is governed by equation (2.27), and thus we can write

$$\frac{\bar{\tau}_c}{\bar{\tau}_e} = g \exp\left(\frac{E_T - E_F}{kT}\right). \tag{3.1}$$

Figure 11 in conjunction with equation (3.1) allows one to identify the times in the up state with electron capture and the down times with emission. On electron capture into a localized electronic state, it would appear that the negative electrostatic potential set up by the trapped charge is responsible for a localized increase in channel resistance. We shall comment on this in greater detail below (section 3.2.3).

In section 2.1 we proved that the up and down times of an RTS are exponentially distributed, and that the mean of the distribution is equal to its standard deviation.

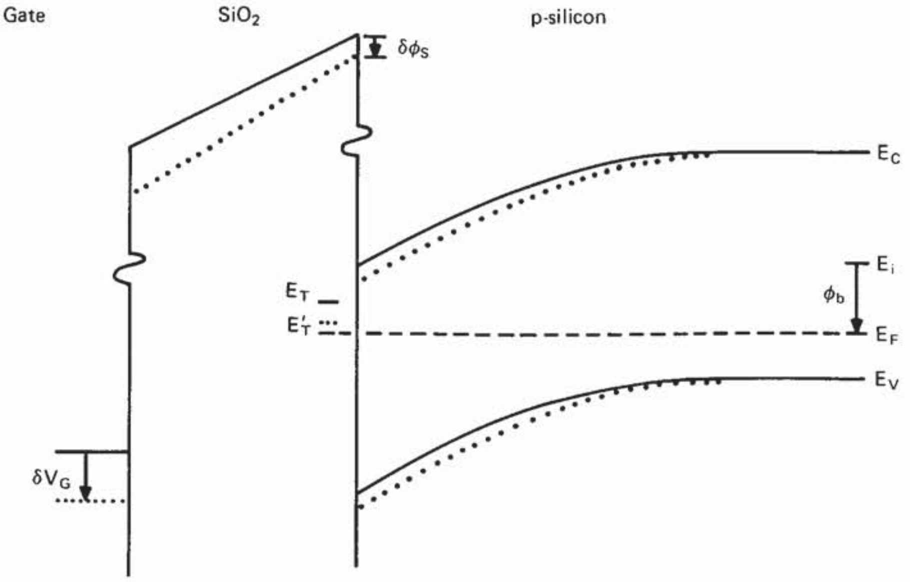


Figure 12. Band bending in n-channel silicon MOSFET. The dotted lines show the changes accompanying a positive increment in gate voltage δV_G. $\delta\phi_s$ is the change in surface potential. E_T and E_T' denote the trap energy-level positions before and after changing gate voltage. ϕ_b denotes the potential of the bulk Fermi level E_F with respect to the intrinsic level E_i.

Figure 13 shows the measured distributions of down (emission) times from a single RTS. For this particular RTS the distribution of up (capture) times was virtually identical. The capture and emission times were exponentially distributed, and thus the switching is governed by the two attempt rates $1/\bar{\tau}_c$ and $1/\bar{\tau}_e$. The means and standard deviations were also in close agreement. A slight departure of the data from the theoretical curve at very small times is expected owing to the limited resolution of the experimental sampling rate.

Let us consider the consequences of the RTSs shown in figures 1 and 11 being due to multi-electron capture into a single defect rather than just single-electron capture. For the purposes of simplicity we shall consider two-electron capture, although this can obviously be generalized. On this basis, the time in the high-current state corresponds to the (rate-limiting) capture of the first electron. This is then followed by the (unseen) very fast capture of the second electron. Thus the time spent in the one-electron level is below the experimental resolution limit.

We shall denote the high-current level as level 0, the (unseen) middle level as level 1, and the low-current level as level 2. Then, using equation (2.22), we find the following relative probabilities of occupation (all degeneracies are taken to be equal):

$$\frac{p(1)}{p(0)} = \exp\left[-\frac{E(1/0) - E_F}{kT}\right], \tag{3.2 a}$$

$$\frac{p(2)}{p(1)} = \exp\left[-\frac{E(2/1) - E_F}{kT}\right], \tag{3.2 b}$$

$$\frac{p(2)}{p(0)} = \exp\left[-\frac{E(1/0) + E(2/1) - 2E_F}{kT}\right]. \tag{3.2 c}$$

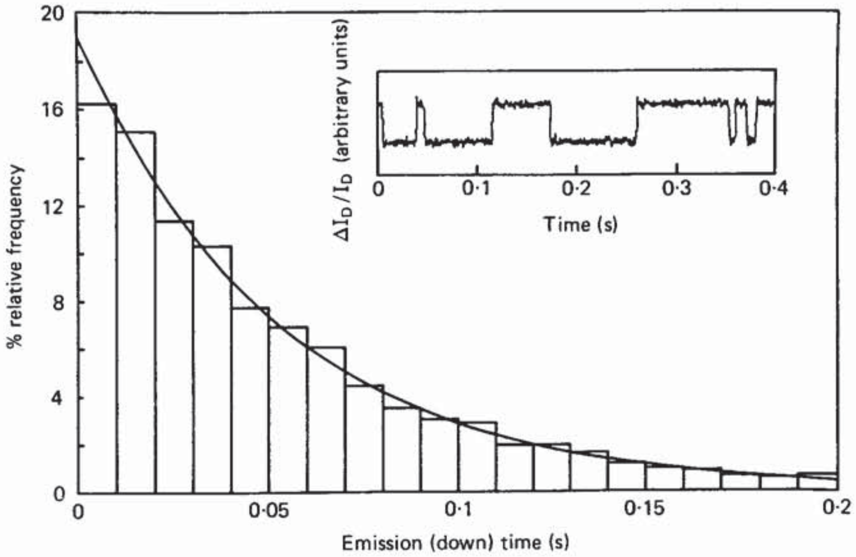


Figure 13. Distribution of 4425 emission times for device H6 at 95 K and $V_G = 1{\cdot}15$ V, showing that the time is distributed exponentially. $\bar{\tau}_e = 0{\cdot}0528$ s, standard deviation 0·0505 s. The inset shows a portion of the *I–t* characteristic; the down time corresponds to emission. From Kirton *et al.* (1989a).

Since $p(1) \approx 0$, this implies that $E(1/0) - E_F$ is large and positive and $E(2/1) - E_F$ is large and negative with respect to kT. Further, since $p(2) \approx p(0)$, this requires $E(1/0)$ and $E(2/1)$ to be roughly equidistant from the Fermi level. Thus the occupancy level $E(1/0)$ must be $\geqslant 10\,kT$ above the Fermi level and $E(2/1) \geqslant 10\,kT$ below. Overall, the defect must exhibit very strong negative-U properties (Anderson 1975). Such a scenario is by no means impossible, but let us now investigate how consistent this model is with further interpretation of the data.

Following Ralls *et al.* (1984), we shall now outline how the behaviour of the mark-space ratio of the RTS can be used to estimate the distance of the trap into the oxide. Taking the logarithm of both sides of equation (3.1) and differentiating with respect to gate voltage, we find for the one-electron case

$$\frac{\mathrm{d}}{\mathrm{d}V_G}(\Delta E_{TF}) = \frac{kT}{q}\frac{\mathrm{d}}{\mathrm{d}V_G}(\ln\bar{\tau}_c - \ln\bar{\tau}_e) \quad (\mathrm{eV\,V^{-1}}), \tag{3.3}$$

where $\Delta E_{TF} = E_T - E_F$. Thus, for a given increment in gate voltage, the measured changes in $\bar{\tau}_c$ and $\bar{\tau}_e$ allow one to estimate the change in separation of the trap energy level and Fermi level: see figure 12. The change in surface potential $\delta\phi_s$ can be estimated from standard MOSFET analysis (Pao and Sah 1965). The distance of the trap into the oxide is then given by the relation

$$\frac{q(\delta V_G - \delta\phi_s)}{t_{ox}} = \frac{\delta(\Delta E_{TF}) - q\,\delta\phi_s}{d}, \tag{3.4}$$

where t_{ox} is the thickness of the gate oxide. Ralls *et al.* found values of d up to 2 nm. We extended this work (Kirton *et al.* 1989a) and showed that the gate-voltage dependence of the carrier trapping times incorporates quite subtle effects, which can give rise to anomalous estimates of the distance of the traps from the interface when the device is operating around threshold. This is discussed in some detail in section 5.1.4.

If the capture process involves two-electron capture then

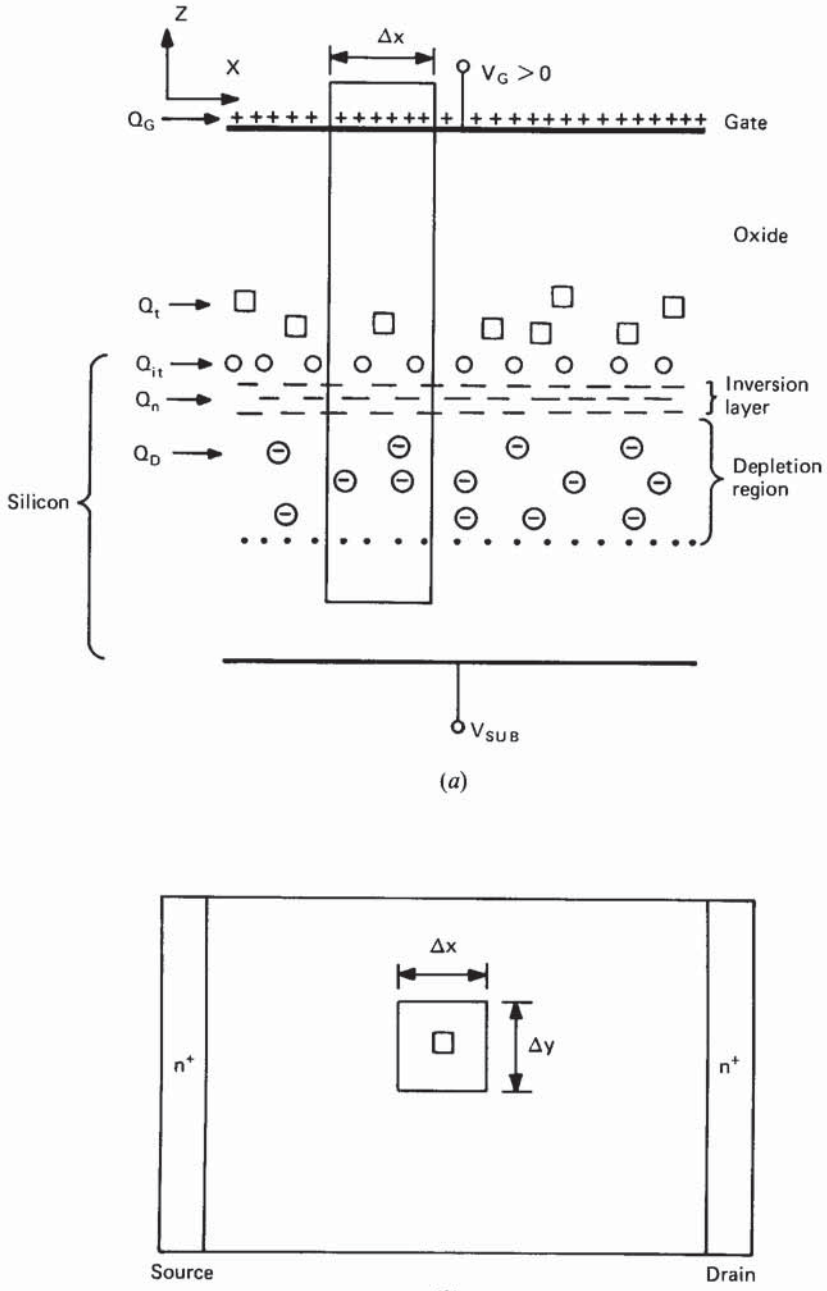
$$\frac{\mathrm{d}}{\mathrm{d}V_{\mathrm{G}}}(2\,\Delta E'_{\mathrm{TF}}) = \frac{kT}{q}\frac{\mathrm{d}}{\mathrm{d}V_{\mathrm{G}}}(\ln\bar{\tau}_{\mathrm{c}} - \ln\bar{\tau}_{\mathrm{e}}) \quad (\mathrm{eV\,V^{-1}}), \tag{3.5}$$

where $\Delta E'_{\mathrm{TF}} = \frac{1}{2}[E(1/0) + E(2/1)] - E_{\mathrm{F}}$. Using equation (3.5) for devices operating in strong inversion, we find the potential change at the trap is usually about half the change at the surface. Since the potential at the inversion-layer charge centroid moves at half the rate of the surface potential (Brews 1978), this places the trap in the middle of the inversion layer, that is, in the silicon rather than the oxide. For the weak-inversion traps that we have studied, namely G8 (table 3) and SP5 (section 6.6), we find that two-electron trapping places the traps 150 and 85 nm into the bulk silicon respectively. (Single-electron trapping places the defects in the oxide.) For G8, this gives a cross-section for capture of the first electron $\sim 10^{-17}\,\mathrm{cm}^2$ and for the second electron $\sim 10^{-13}\,\mathrm{cm}^2$.

So if these telegraph signals are due to multi-electron trapping, then they represent Si rather than interface or $\mathrm{SiO_2}$ defects, but in practice a wide range of (model-independent) activation energies for capture are measured (table 3). Whereas this observation is consistent with capture into defects in the amorphous oxide, for deep levels in the bulk that exhibit very strong negative-U properties one might expect a few well defined energies corresponding to defects of particular chemical and structural origin. Moreover, as we demonstrate in section 7, on summing noise measurements made on small devices one finds very good agreement with the $1/f$ power spectra measured in large devices (figure 53). Since $1/f$ noise requires a distribution of time constants (cross-sections) spanning many decades, it is clear that the majority of the trapping must take place into oxide defects. Typically, the presence of bulk silicon traps shows up as a distinct Lorentzian on a $1/f$ background. We have found no evidence of this effect (figure 17). Overall, single-electron trapping into defect states in the oxide provides the simplest explanation for the majority of the data; though this cannot rule out the possibility that a small proportion of the defects are in fact multi-electron silicon traps. Indeed, a small proportion of the RTSs shows complex behaviour (section 6), and in DLTS experiments transients have been observed that are most simply explained as multi-carrier emission (Karwath and Schulz 1989). We pursue this matter further in section 3.2.3.

3.2.3. *The amplitude of RTSs*

With reference to figure 11, we shall take our initial working hypothesis to be the following: after an electron is trapped into an $\mathrm{Si/SiO_2}$ interface defect state, the reduction in source–drain current comes about through a reduction in the number of free carriers in the channel. The strong-inversion case is simplest to consider first. In this regime the screening of the trapped charge is carried out by the inversion-layer electrons. We therefore expect on electron capture a reduction of unity in the total number of free carriers in the inversion layer, $N_{\mathrm{inv,tot}}$, and hence (ignoring scattering) $\Delta I_{\mathrm{D}}/I_{\mathrm{D}} = 1/N_{\mathrm{inv,tot}}$. As the gate voltage is reduced to threshold and below, the screening of the trapped charge is now principally carried out by the depletion region and the gate; the estimation of the reduction in total carrier number becomes accordingly



(*b*)

Figure 14. (*a*) Charge structure in a MOSFET. Q_t, Q_{it}, Q_n, Q_D and Q_G are the charge densities (per unit area) associated with oxide traps, interface states, the inversion layer, the depletion region and the gate respectively. (*b*) Plan view showing small area of channel cored out by electron trapped in oxide defect. After Reimbold (1984).

more complex. We shall now outline a simple theory for the expected variation in $\Delta I_D/I_D$ from weak to strong inversion.

Following Reimbold (1984), figure 14 (*a*) depicts the location of charge in an MOS structure. We define Q_G, Q_{it}, Q_D, Q_n and Q_t to be the charge density (per unit area) associated with the gate, interface traps, depletion region, inversion-layer and the oxide traps respectively. When the trapped charge Q_t fluctuates at fixed V_G the charge conservation in the structure is such that

$$\delta Q_G + \delta Q_{it} + \delta Q_D + \delta Q_n + \delta Q_t = 0. \tag{3.6}$$

These fluctuations can be related to the change in surface potential $\delta\phi_s$ via the following relations:

$$\delta Q_G = -C_{ox}\,\delta\phi_s, \tag{3.7 a}$$

$$\delta Q_{it} = -C_{it}\,\delta\phi_s, \tag{3.7 b}$$

$$\delta Q_D = -C_D\,\delta\phi_s, \tag{3.7 c}$$

$$\delta Q_n = -C_n\,\delta\phi_s. \tag{3.7 d}$$

C_{ox}, C_{it}, C_D and C_n are the capacitances (per unit area) associated with the oxide, interface traps, depletion region and channel respectively. Now

$$Q_n = Q_0 \exp\left(\frac{q\alpha\phi_s}{kT}\right), \tag{3.8 a}$$

and therefore

$$\delta Q_n/\delta\phi_s = \beta\alpha Q_n, \tag{3.8 b}$$

where $\beta = q/kT$, Q_0 is a constant, and α has a value of 1 in weak inversion falling to 0·5 in strong inversion (Brews 1978).

Equations (3.6) and (3.7 *a–d*) allow us to write

$$\delta Q_t/\delta\phi_s = C_{ox} + C_{it} + C_D + C_n. \tag{3.9}$$

Using the relation $\delta Q_n/\delta Q_t = (\delta Q_n/\delta\phi_s)\,\delta\phi_s/\delta Q_t$ and equations (3.7 *d*) and (3.9), we obtain

$$\frac{\delta Q_n}{\delta Q_t} = \frac{\delta Q_n/\delta\phi_s}{C_{ox} + C_{it} + C_D - \delta Q_n/\delta\phi_s}. \tag{3.10}$$

In the limit of strong inversion the terms in the denominator are dominated by $\delta Q_n/\delta\phi_s$, and we find $\delta Q_n/\delta Q_t = -1$; if the trap gains one electron then the inversion layer loses one electron, and *vice versa*. As weak inversion is approached $|\delta Q_n/\delta Q_t| < 1$, corresponding to charge sharing between the gate and the inversion and depletion layers.

If we assume that all changes in the charge distributions on electron capture into an oxide defect are located within the small area $\Delta a = \Delta X\,\Delta Y$ as shown in figure 14 (*b*) then we can write

$$\Delta a\,\delta Q_t = -q, \tag{3.11}$$

a single electronic charge. Further, we assume that within this area Δa

$$\delta Q_n = -Q_n. \tag{3.12}$$

Thus the model is based on total exclusion of inversion-layer charge. In practice, the charge does not fall to zero. Integrating Poisson's equation with a trapped charge at the Si/SiO$_2$ interface along a line perpendicular to the surface above the trap, we found gave a reduction to 22% of the value without the trap for the device biased in weak inversion. Thus our model is oversimplified and is not realistic, but it is nevertheless a useful first approximation. We estimate Δa as follows. Equations (3.8 *b*) and (3.10) can be combined to give

$$\frac{\delta Q_n}{Q_n} = \frac{\alpha\beta\,\delta Q_t}{C_{ox} + C_{it} + C_D - \alpha\beta Q_n}. \tag{3.13}$$

Multiplying the top and bottom of the right-hand side of equation (3.13) by area Δa and confining all changes in charge density to this area, as well as noting the relationships (3.11) and (3.12), we find

$$\Delta a = \frac{\alpha\beta q}{C_{ox} + C_{it} + C_D - \alpha\beta Q_n}. \tag{3.14}$$

Since we have assumed that within area Δa all inversion-layer charge is excluded (i.e. its resistance is infinite) it is a straightforward matter to show that

$$\frac{\Delta I_D}{I_D} = \frac{\Delta R}{R} = \frac{\Delta a}{A} = \frac{\alpha\beta q}{A(C_{ox} + C_{it} + C_D - \alpha\beta Q_n)}, \tag{3.15}$$

where R is the channel resistance and A is the device area.

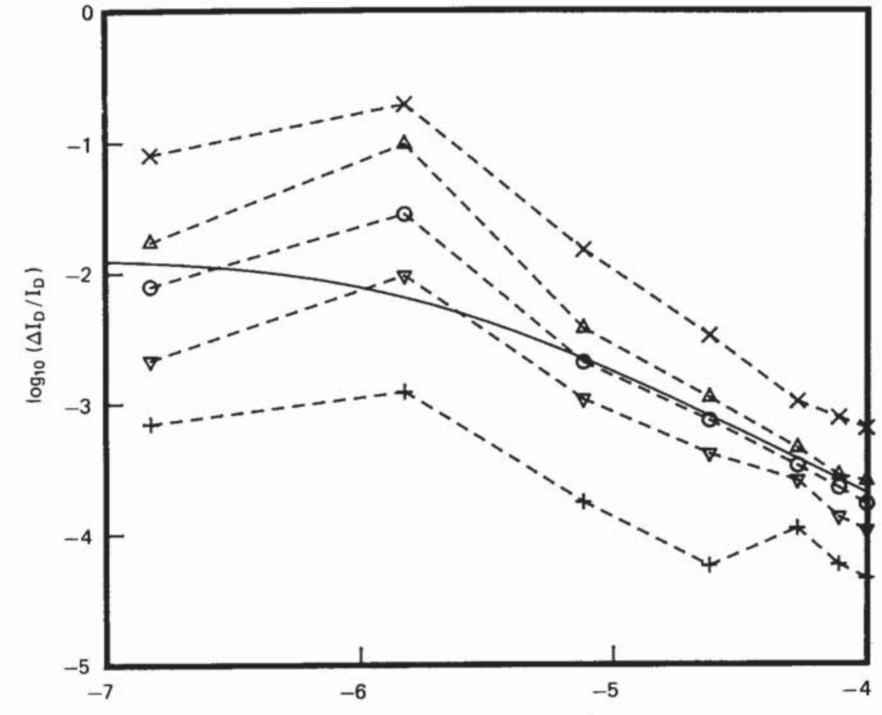


Figure 15. Measured values of RTS amplitudes recovered from a survey of small-area ($0{\cdot}375\,\mu m^2$) MOSFETs, from the same wafer, as a function of channel conductivity $G = I_D l / V_D w$ (Siemens/□). The dashed lines show the behaviour of the 0, 25, 50, 75 and 100th percentiles (see the text). The solid line shows the theoretical value of $\Delta I_D / I_D$ obtained from equation (3.15).

In figure 15 we show the measured values of $\Delta I_D/I_D$ recovered from a survey of small-area ($0{\cdot}375\,\mu m^2$) MOSFETs from the same wafer. The data were obtained by choosing seven values of the source–drain current (at fixed V_D) and at each value the amplitudes of the RTSs that were visible were measured. The time window in which the measurements were carried out was 10^{-3}–10^2 s. At each current the number of distinct traps seen was 51, 30, 70, 89, 56, 60 and 58 in 27, 12, 28, 18, 11, 10 and 12 devices, in order of increasing current magnitude. For a given current, the distribution of amplitudes was sorted into percentiles. For example, the 50th percentile is the value below which 50% of the distribution lies. On figure 15 we have marked the 0, 25, 50, 75 and 100th percentiles. The theoretical values of $\Delta I_D/I_D$ from equation (3.15) have also been plotted. Apart from the data points around $10^{-6}\,S\,\square^{-1}$—which represent only 30 traps—the theoretical curve essentially passes through the 50th percentile of the data.

The data presented in figure 15 suggest that the average behaviour is reasonably represented by our simple theory. However, it is somewhat disconcerting to see that a large number of RTS amplitudes appear to correspond to significantly less than or significantly greater than one electron trapped, as was first noted by Ralls *et al.* 1984. In addition, Schulz and co-workers found that single RTS amplitudes when followed over a range of channel conductivities sometimes showed quite complex behaviour departing significantly from the form given by equation (3.15) (Kirton *et al.* 1989b).

In section 3.2.2 we found that the conditions necessary for multi-electron trapping at a single defect were such that it was an unlikely event. In strong inversion the inversion-layer screens the trapped charge, so that the area of the channel affected when two electrons are trapped, for example, is just double that when one electron is captured. A scatter plot of the amplitude data (figure 16) shows that in the strong-inversion regime there is no evidence of integer multiples of one electron being trapped. All that is observed is a near-continuous distribution with a few larger amplitude traps. We have found that some of the very smallest amplitude traps of figure 15 showed a much weaker dependence on gate voltage than normal. This suggests (Karwath and Schulz 1989, Jantsch and Kircher 1989) that they reside in the thicker oxide of the 'bird's beak' along the edge of the channel (figure 10 (*c*)).

Very recently, Restle (1988) reported on work which provides strong evidence that the RTSs are due to defects residing in the channel. By investigating the effects of changing the source-drain voltage, he was able to locate the position of individual defects between source and drain. Restle was also able to monitor the effect on the RTS amplitude of sweeping the pinch-off region past a trap. In section 7.1.1 we shall discuss the results of summing noise power measurements made on many small devices and comparing the result with measurements made on large devices (figure 53). The results provide evidence that the measured distribution of amplitudes in the small devices is not being grossly distorted by any inhomogeneities specific to small devices or by the proximity of the device perimeter and hence is reasonably characteristic of the channel.

It is well known that there will be fluctuations in the surface potential due to a spatially random distribution of fixed charge near the Si/SiO$_2$ interface. Brews (1975a, b) has developed a perturbation theory that predicts that the device mobility will be reduced and take the form $\mu \approx \mu_0(1 - \frac{1}{2}\langle\sigma_s^2\rangle)$, where σ_s is the standard deviation of the surface-potential fluctuations. In weak inversion this can lead to inhomogeneous transport and a measurable reduction in mobility (Muls *et al.* 1978); however, for modern devices with low fixed charge there is little effect (Ando *et al.*

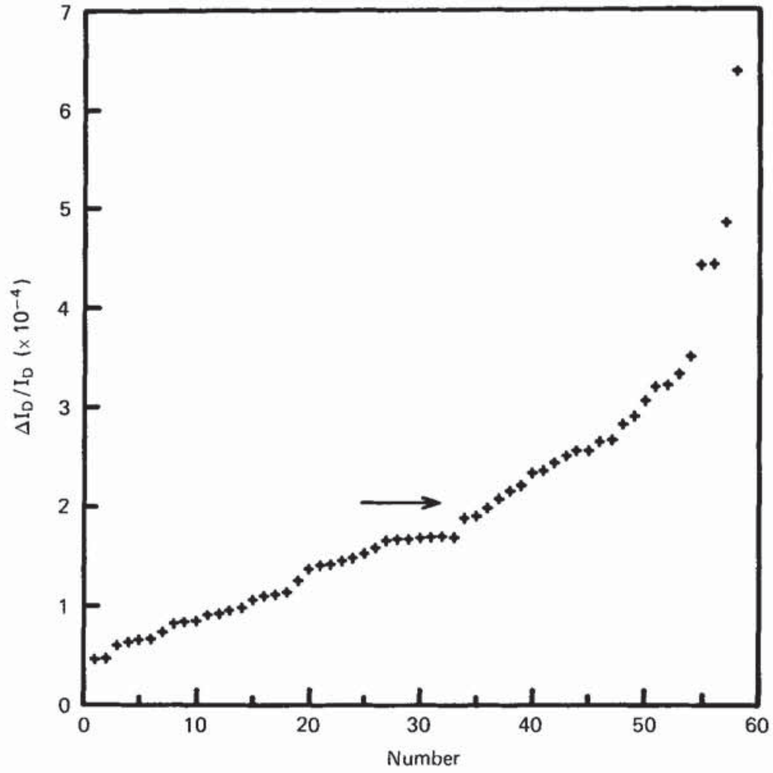


Figure 16. Scatter plot of the RTS amplitude data corresponding to the value of $G = 10^{-4}\,\mathrm{S}\,\square^{-1}$ on figure 15. $V_G = 3{\cdot}8\,\mathrm{V}$, $T = 293\,\mathrm{K}$, $V_D = 100\,\mathrm{mV}$, $I_D = 6{\cdot}7\,\mu\mathrm{A}$.

1982, p. 493). This sort of mechanism could be used to explain the wide range of amplitudes in weak inversion. Above threshold, as the carrier concentration is increased, $\langle\sigma_s^2\rangle$ tends rapidly to zero as the fluctuations are screened out. Any explanation based on this effect in strong inversion, where we still observe a range of amplitudes, would have to argue that the trapping states are physically associated with the fixed charge.

It therefore appears that the wide distribution of amplitudes is a real effect and is not fully accounted for by edges, multi-electron capture or potential fluctuations. Thus we can pose the following question: By what means does a trapped electron modulate the conductivity of the inversion layer? We have already seen that the average effect is accounted for by a reduction in number of carriers in the channel. Is it possible that some additional factor, such as a change in carrier mobility, is also involved? In essence, we are attempting to understand the effect of a single point scatterer on the conductivity (Landauer 1957, 1975). As yet, a theory explaining the data of figure 15 is not available, and the amplitude of RTSs in microstructures remains one of the outstanding fundamental problems to be addressed. In the remainder of this section we shall probe this and related areas.

Careful inspection of figure 6 for RTSs in a MOSFET at cryogenic temperatures provides evidence that carrier mobility is affected by a trapped electron. We see that the high-resistance state corresponds to electron capture and the low-resistance state to electron emission. Thus for this particular RTS the channel resistance is *reduced* when an electron is trapped, which is in contrast with all our measurements around room temperature. Ralls *et al.* (1984) found examples of both types of polarity, namely traps in which the current was reduced and traps in which the current was increased on electron capture. They also noted that the amplitude varied greatly in size and could be greater or smaller than the fractional change due to the removal of a single channel electron.

In the low-temperature regime investigated by Ralls and co-workers it is Coulombic (including interface-roughness) scattering that predominantly determines the carrier mobility. One can thus envisage the following two scenarios: a positively charged scattering centre is neutralized by electron capture and is thus turned off, corresponding to a discrete increase in current; and a neutral centre becomes singly negatively charged, giving a reduction in current due to increased scattering. Ralls *et al.* suggested that the range of amplitudes may be accounted for by some scattering centres being more strategically located than others.

At low temperatures universal conductance fluctuations (UCFs) become important (Lee and Stone 1985, Feng *et al.* 1986, Skocpol *et al.* 1986). UCFs arise in the regime in which the elastic scattering length is much smaller than the sample length L, and the inelastic scattering length is larger than L. One then finds a large random component of the conductance (of order e^2/h), which depends on the detailed relative positions of the elastic scattering sites. The random component can be measured as a sample-to-sample variation in the conductance, or within a single sample as a function of magnetic field since the interference terms in the scattering also depend on magnetic field. Feng *et al.* (1986) pointed out that the measured conductance should be sensitive to the motion of a single atom, which changes the impurity configuration. Beutler *et al.* (1987) carried out measurements on the temporal behaviour of the conductance in thin bismuth wires and films. They observed discrete switching in the conductance, which they ascribed to single impurity motion. Very recently, Birge *et al.* (1989) have reported UCF $1/f$ noise in bismuth due to the motion of many defects. Such phenomena may provide an explanation for the wide range of amplitudes, and even change of sign, seen by Ralls *et al.* Clearly though, RTSs in MOSFETs at room temperature are quite distinct from any UCF. Although the scattering rates in the silicon inversion layer are not known accurately, at elevated temperatures phonon scattering (which is phase-breaking) dominates over elastic Coulombic scattering.

As we shall discuss in section 6.5, low-frequency noise in metals is correlated with the presence of complex defects. Ralls and Buhrman's (1988) recent findings on the noise in metallic microstructures provide striking evidence that by changing its configuration a defect can alter the sample resistivity via changes in the scattering cross-section (Pelz and Clarke 1987). Is it then possible that such changes in configuration, with no change in charge state, for Si/SiO_2 interface defects could modulate the channel conductivity? In section 6.6 we shall provide evidence that (complex) oxide defects may be able to do just that. The spread in the data of figure 15 could then be explained by some defect structures being more efficient at altering inversion-layer conductivities than others. The physical mechanism by which this comes about is still uncertain. Clearly this is an area in which much fundamental work remains to be done.

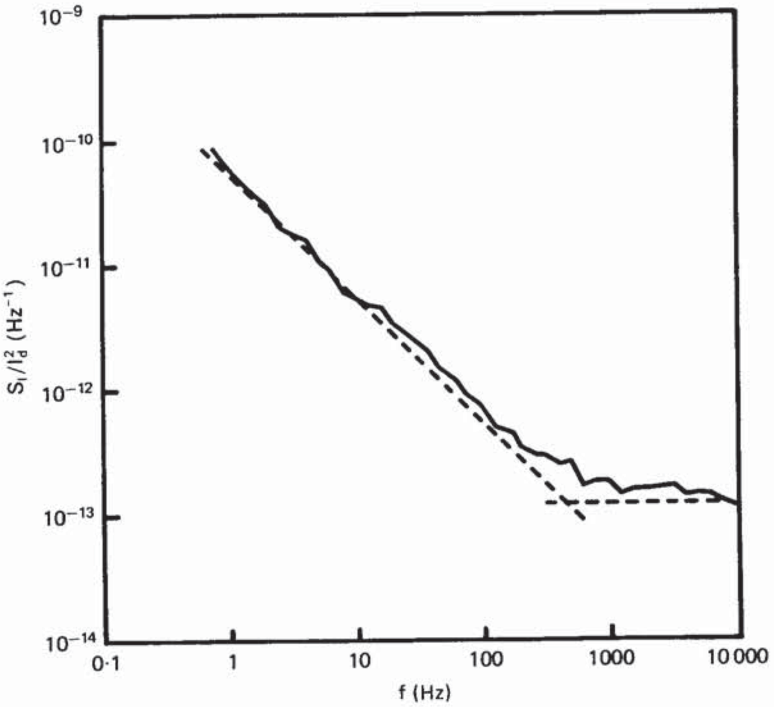


Figure 17. Noise power spectrum measured in large-area commercial MOSFET at room temperature. At low frequencies the slope is close to -1, corresponding to $1/f$ noise. At higher frequencies (> 500 Hz) the white noise in the channel dominates. $w/l = 20\,\mu\text{m}/20\,\mu\text{m}$, $V_D = 100$ mV, $V_G = 1{\cdot}5$ V, $t_{ox} = 40$ nm.

4. 1/*f* noise in microstructures

Figure 17 shows a typical example of the noise power spectral density measured in a commercial, large-area (20 μm × 20 μm) MOSFET. On a log–log plot the 1/*f* spectrum appears as a straight line with a slope close to -1. Now we have already seen that in small MOSFETs the current fluctuations appear as random telegraph signals (RTSs) (in section 2.3 we showed that the power spectrum of an RTS is Lorentzian) so that the fundamental question that we now wish to address is the following: What is the relationship between the RTSs observed in small devices and the 1/*f* spectrum measured in large devices? We shall see that the answer to this question is quite straightforward: RTSs are a product of the decomposition of the 1/*f* spectrum into its individual fluctuating components; conversely the 1/*f* spectrum can be viewed as a superposition of individual trapping events, each generating an RTS in the time domain.

The defect origins of 1/*f* noise have long been suspected, but not definitely proved to the satisfaction of all. On the basis of the work of McWhorter (1957), Ralls *et al.* (1984) suggested that RTSs were probably the cause of 1/*f* noise in MOSFETs (figure 6), but they were unable to demonstrate conclusively the full superposition process to give a 1/*f* spectrum. This was due to the fact that in the very small devices which they were studying (about 0·1 μm^2) it was not possible, even at elevated temperatures, to observe the fluctuations of more than a handful of traps at any particular gate voltage. Studies by Rogers and Buhrman (1984) on small-area metal–insulator–metal (MIM) tunnel junctions, Restle *et al.* (1985) on small-area silicon-on-sapphire resistors and Uren *et al.* (1985) on a range of MOSFETs of different areas provided unambiguous evidence that $1/f$ noise in these systems is generated through the fluctuation in occupancy of individual defect states modulating the conductivity. We shall now consider these investigations in turn.

4.1. *$1/f$ noise in small-area metal–insulator–metal tunnel junctions*

The devices studied by Rogers and Buhrman (1984) were Nb–Nb_2O_5–PbBi tunnel junctions with active areas spanning the range 5×10^{-2}–$1{\cdot}0\,\mu m^2$. Current flowed between the two metal contacts via direct tunnelling through the band gap of the insulator. Defect levels present in the insulator and situated in the band gap were able to capture the tunnelling electrons. Schmidlin (1966) showed that charged ions in the insulator can have a significant effect on barrier shape. In sufficiently small devices with only a few active traps, the fluctuating occupancy of an individual trap has a measurable effect on the current through the device.

In figure 18 plots of $fS_V(f)$ against frequency obtained by Rogers and Buhrman are displayed. A power spectrum exhibiting $1/f$ behaviour would appear as a horizontal trace on this modified plot. Despite the rolling behaviour of the spectra, they display

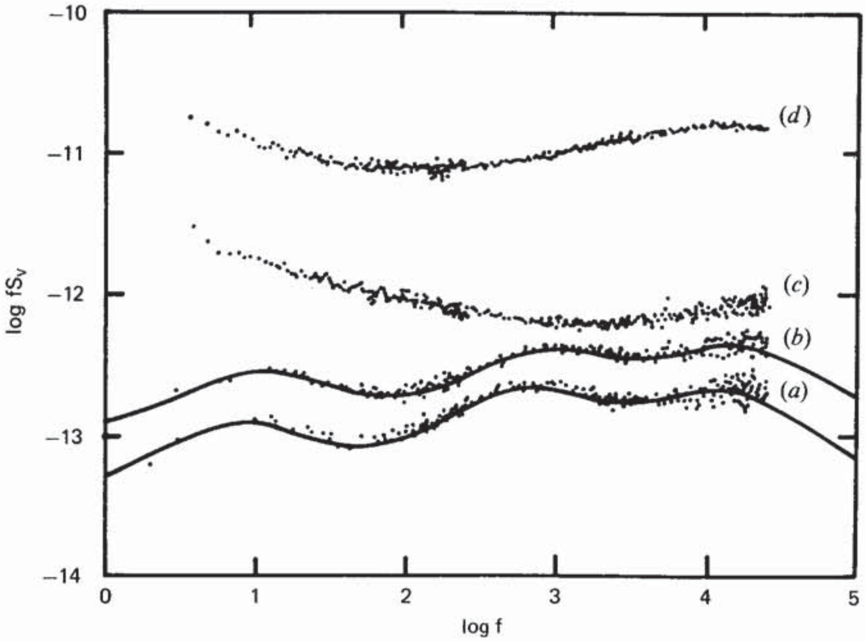


Figure 18. $fS_V(f)$ for a MIM tunnel junction with $A = 10^{-9}$ cm^2 and $R = 110\,\Omega$. Traces (*a*) and (*b*) are at 24 K and 65 μA and 105 μA respectively. The solid lines are nonlinear least-squares fits using a three-Lorentzian function plus a small $1/f$ term. Traces (*c*) and (*d*) are at 85 μA and 170 K and 260 K respectively. They cannot be fitted with less than five Lorentzians. After Rogers and Buhrman (1984). © American Physical Society. Reproduced with permission.

a trend about a flat $1/f$ behaviour. In addition, Rogers and Buhrman found the magnitude of the spectra to scale with I^2 and $1/A$, where A is the device area. We shall now show that this behaviour is consistent with a model of individual fluctuators affecting the conductance locally over a small area, independent of the full area of the device.

Consider a device that one can imagine to have been partitioned into individual units: the conductance of each unit is locally modulated only by the trapping centre contained within it. Now the power spectrum of the current fluctuations through the device can be written as (see equation (2.19 *a*))

$$\frac{S_I(f)}{I^2} \propto \sum_{k=1}^{N_{\mathrm{dev}}} \frac{(\Delta I/I)_k^2}{(\bar{\tau}_0 + \bar{\tau}_1)_k[(1/\bar{\tau}_0 + 1/\bar{\tau}_1)_k^2 + (2\pi f)^2]}, \tag{4.1}$$

where the summation is over all traps ($k = 1, 2, \ldots, N_{\mathrm{dev}}$) contained within the device. For a uniform current flow I the ratio $(\Delta I/I)_k \propto 1/A$, where A is the device area, and for a constant trap density $N_{\mathrm{dev}} \propto A$. Thus if we assume that N_{dev} is always large enough to ensure that the full distribution of time constants is summed over then we should expect to find the following continuous and smooth relationship:

$$\frac{S_I(f)}{I^2} = \frac{S_V(f)}{V^2} \propto \frac{1}{A}. \tag{4.2}$$

In very small devices where the number N_{dev} is small we should expect irregular lumps and bumps to appear in the spectrum. In this respect, a most important feature observed by Rogers and Buhrman was that at low temperatures the observed spectra could be accurately and uniquely fitted with a finite number of Lorentzian spectra plus a small (of order 1–10%) residual $1/f$ component. In addition, every tunnel junction measured had its own particular collection of Lorentzians at a particular temperature and applied voltage. The overall noise level was relatively constant from device to device, at the level predicted from the $1/A$ dependence of equation (4.2). This sample-to-sample variation is a reflection of the limited number of active traps per device. On raising the temperature, the number of active defects increased with increasing thermal energy. Between 100 and 150 K the spectra became difficult to fit unambiguously. Overall, Rogers and Buhrman observed a smooth transition from nearly classic $1/f$ noise at high temperatures to a spectrum composed of a unique superposition of distinct Lorentzian components for temperatures $\leqslant 80$ K.

4.2. $1/f$ *noise in small silicon-on-sapphire resistors*

An elegant study of both the statistics and area dependence of $1/f$ noise in silicon-on-sapphire (SOS) resistors was made by Restle *et al.* (1985). Essentially, the work had two main goals. First, through a coupled theoretical and experimental approach, they wished to distinguish between two competing views on the origins of $1/f$ noise: the so-called two-state model (i.e. superposition of RTSs) and the random walk in a random potential (RWRP) model (i.e. slow reconfigurations of a glass). Secondly, through statistical tests via covariance matrices, they aimed to test for non-Gaussian behaviour in the noise. This latter aspect of the work will be discussed in section 6, when we consider the phenomenon of complex telegraph noise.

Restle *et al.* used a Monte Carlo procedure to generate theoretical $1/f$ noise spectra. The spectra were composed by summing the fluctuations of 30 two-state (RTS) systems with different characteristic frequencies and duty cycles (i.e. ratios of up to down times). Figure 19 depicts four such spectra plotted as noise per octave