Hot Carrier Effects in IGFETs

Hercules Pereira Neves

Ph.D. Thesis

Department of Electrical Engineering
University of Edinburgh
1991

EX BIBL. UNIV. EDINBURGH

DECLARATION

I hereby declare that I alone have composed this thesis and that, except where stated otherwise, the work presented within is my own.

Hercules Pereira Neves

Aos meus dois melhores amigos: Élcio Teixeira Neves (meu pai) e Lúcia Maria Pereira Neves (minha mãe).

Acknowledgements

I am very grateful to my supervisor, Professor John M. Robertson, for his cooperation even before my arrival in Scotland. His guidance throughout the development of my work, his availability for discussion on all matters and the autonomy he conferred to my decisions undoubtedly set a paradigm of postgraduate supervision.

I would also like to thank my second supervisor, Dr. Anthony J. Walton, whose advice was particularly useful at the beginning of my research work.

Many thanks to Dr. Robert J. Holwill, Director of the Edinburgh Microfabrication Facility, who granted me access to personally operate all equipment in the Facility and gave priority to my fabrication batches when required. I am also grateful to all of those who helped and advised me on fabrication matters; particularly to Mr. Alan M. Gundlach (EMF Operations Manager), Dr. J. M. T. Stevenson, Mr. J. Gow and Mr. Z. Chen. I am also grateful to all the EMF staff for their kind assistance and friendship.

I would also like to express my gratitude to Dr. J. J. Delima and Mr. R. Hannah for reviewing part of my text.

I am grateful to my friends Catherine McGuigan, Vince Duffy, Robbie Hannah, Alan Moir, Martin Fallon, Bob Holwill, Alan Ross and the Margutti family who made my stay in Scotland even more pleasant.

I am deeply thankful to CNPq (Conselho Nacional de Desenvolvimento Cientifico e Tecnologico) for trusting me and awarding me a scholarship, without which my postgraduate studies would be impossible.

Finally, I would like to thank my parents for their patience in looking after my affairs in Brazil during my absence.

ABSTRACT

In this work the analysis of hot carrier generation and injection in insulated-gate field effect transistors (IGFETs) is carried out with the aid of offset gate transistors. Several sets of transistors with systematically incremented gate offsets were produced for this investigation. The inclusion of photolithographically and angled-implant defined LDD (lightly doped drain) regions was also attempted, but no consistent calibration technique was found for the former and fabrication problems made the latter impractical. The charge pumping technique and the gated diode current method were employed in the evaluation of the generation of interface states and fixed charges. An adapted double-pulse charge pumping method was used to assess the energy distribution of interface states across the bandgap. The use of offset gate transistors made it possible to identify hot electron and hot hole degradation separately under different biasing conditions. Additionally, it suggested a clearer interpretation of the possible mechanisms involved in the degradation process.

Table of Contents

Chapter 1 - Introduction 1

Chapter 2 - Hot Carriers in Silicon 8

2.1. Introduction 8

2.2. Macroscopic approach 8

2.3. Carrier Statistics 9

2.4. Hot electron generation 12

2.5. The Boltzmann transport equation 17

2.6. The Monte Carlo technique 19

References 24

Chapter 3 - Hot Carrier Investigation in IGFET Devices 26

3.1. Introduction 26

3.2. Hot carrier injection into SiO_2 27

3.3. IGFET biasing conditions for injection 31

3.4. Experimental methods 40

3.4.1. The charge pumping method 41

3.4.2. The gated diode method 50

References 53

Chapter 4 - Device Fabrication 56

4.1. Introduction 56

4.2. First fabrication 58

4.2.1. Design considerations 58

4.2.2. Fabrication process 63

4.2.2.1. Alignment control 63
4.2.2.2. Analysis of the contact implant 84
4.2.2.3. Polysilicon recrystallisation 88
4.3. Second fabrication 88
4.3.1. Design considerations 90
4.3.2. Fabrication process 90
4.3.2.1. Alignment control 101
4.3.2.2. Narrow line photolithography 101
4.3.2.3. Implant barrier fabrication 107
4.4. Conclusion 121
References 122
Chapter 5 - Device Testing 123
5.1. Introduction 123
5.2. Initial assessment 123
5.3. Description of the measurement system 124
5.4. LDD verification 125
5.5. Gate offset verification 133
5.6. Implementation of the charge pumping method 148
5.7. Implementation of the gated diode method 151
5.8. Implementation of the measurement of interface state energy distribution 154
5.9. Degradation assessment strategy 161
5.10. Experimental results 163
References 189
Chapter 6 - Analysis of Results 192
6.1. Introduction 192
6.2. Significance of the offset gate experiment 192

6.2.1. Case 1: $V_{GS} = V_{DS} = 5V$ (hot electron degradation) 193
6.2.2. Case 2: $V_{GS} = 1V$, $V_{DS} = 5V$ (hot hole degradation) 200
6.2.3. Case 3: $V_{DS} = 5V$, $V_{GS} = 0$ to 5 V at 100 kHz (a.c. degradation) 204
6.3. Predominance of hot electron versus hot hole degradation 204
6.4. Energy distribution of generated interface states 207
6.5. Summary of conclusions 211
References 212
Chapter 7 - Conclusions 213
7.1. General remarks 213
7.2. Suggestions for future work 214
7.2.1. Device design and fabrication 214
7.2.2. Device testing 215

Chapter 1 - Introduction

The insulated gate field effect transistor (IGFET) is most certainly one of the simplest possible three-terminal semiconductor devices. Compared to the bipolar transistor, it is structurally simpler: since it only requires one diffusion step, the junctions are relatively shallow (therefore allowing good diffusion model fitting) and the isolation between adjacent transistors does not have to use the bulk of the silicon substrate. Its concept is also very simple: a unipolar current is controlled by an electric field. Yet it only became available thirty years after it was first predicted. The reason is simple: despite its structural simplicity, the IGFET involves the properties of semiconductor surfaces and interfaces, which constitutes one of the most complex and still poorly understood subjects of materials science.

Fabrication technology evolved and made the IGFET possible. Germanium, which was used as the semiconducting material in the early days of microelectronics was largely replaced by silicon. This breakthrough, which could only happen after extensive research on the production of ultra-pure monocrystalline silicon, brought the advantage of a much better native oxide. Silicon dioxide presents very good characteristics, both chemically and electrically. In contrast with germanium dioxide, it is not water-soluble; thus the predominantly wet chemical processing at the time would have no adverse effects on SiO_2 layers. The properties of the Si/SiO_2 system continued to be investigated and problems such as the ionic contamination of SiO_2 were identified and greatly reduced. Nevertheless, the system has inherent non-idealities which will never be totally eliminated. The very transition between silicon and silicon dioxide is not as abrupt as one may imagine. The oxidation process, on the other hand, introduces its own imperfections, such as hydrogenated compounds. Additionally, the use of modern processes such as electron-beam lithography, ion implantation, etc. also introduce damage to the surface, therefore complicating the interface problem even further. This non-ideal scenario is where the equally complex phenomenon of hot carrier injection takes place.

As integrated circuits become more complex, they incorporate larger numbers of transistors. In order to attain high operational speeds, it is important to keep the delay introduced by each transistor as low as possible. On the other hand, the increase in circuit complexity also increases the chances for failure. For example, the old MSI (medium size integration) circuits with under a hundred transistors had yields above 90% per wafer, whereas for some of the modern ULSI (ultra large scale integration) circuits, with over a million transistors per chip, a yield of 5% is sometimes considered excellent. It is therefore desirable to produce as many chips per wafer as possible in order to reduce costs. The increase of the number of chips per wafer implies in either increasing the wafer size or decreasing the chip size. The wafer size has indeed increased over the past decades but it is limited by practical aspects, like the wafer fragility, the difficulty in executing certain process steps (e.g. ion implantation and lithography) and the difficulty in maintaining the uniformity of layer thicknesses and impurity concentrations over too large an area.

The need to decrease the chip size and to keep a high operational speed calls for the reduction in the size of transistors. Scaling rules were devised [1-3] so that the operating characteristics could be preserved. However, if the biasing voltages are kept the same, there is an increase in the electric field values as the transistor dimensions are reduced. It is possible to reduce the operating voltages, but this is limited by the built-in potential of the junctions, which cannot be scaled. [4] Different techniques have been used to reduce local peaks in the electric distribution. The two main structures which resulted from such effort are the LDD (lightly doped drain) and DDD (doubly diffused drain),[5-9] which are depicted in Figure 1.1. These structures reduce the peak of the electric field near the drain junction by making the junction more gradual. Although LDDs and DDDs reduce the peak electric field, this simply means that the limits imposed by high fields are stretched a little further, but not eliminated.

High electric fields are responsible for the generation of hot carriers, which in turn can be injected towards the Si/SiO_2 barrier and modify the electric properties of the Si/SiO_2 complex. This will lead to the degradation in the operating characteristics of the transistor. Due to the complexity of the Si/SiO_2 system, the mechanisms involved in this degradation are difficult to determine. Furthermore, the injection of hot carriers is

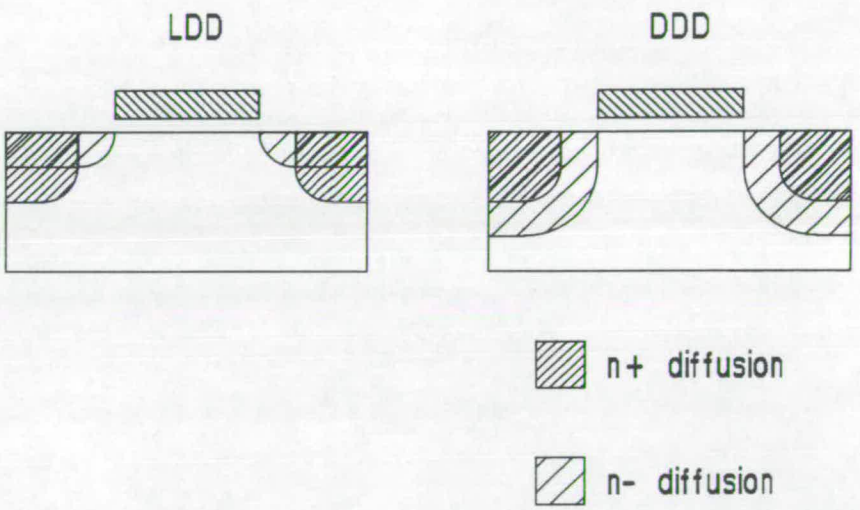


Figure 1.1: Hot-carrier-resistant structures. A decrease in the generation of hot carriers is accomplished by reducing the peak of the electric field near the drain.

difficult to monitor because it requires long periods of time to become noticeable (the number of injected carriers is very small) and also because the injection current is not directly measurable. The distribution of the electric field inside the transistor also complicates the problem because of its two-dimensional character and because carrier heating depends primarily on the field gradient.

Many different techniques have been employed throughout the years for the study of hot carrier degradation in IGFETs. However, apart from the use of hot carrier resistant structures (LDDs, DDDs, etc.), the same transistor structures have been used in virtually all the previous work. This means that hot carrier effects always followed the same generation/injection pattern in all of them. The only significant difference is that smaller transistors will have higher electric fields. All that the different measurement methods could do was to investigate the effects of degradation once they had occurred; any theories regarding the degradation process, which always followed the same pattern, were a subject for speculation. Apart from the post-degradation assessment techniques, only a few variables could be adjusted by the investigator: the length of the transistor channel, the biasing voltages and the temperature. It is evident that the introduction of any new variable would generate a multitude of new information, considering that all the existent assessment methods could be used in conjunction with it. In the present work a new variable, or a new degree of freedom, was introduced: the gate offset. As will be seen in the following chapters, this opens an entirely new perspective to the understanding of hot carrier degradation. This introduction was done in an analytical way, i.e. offsets were introduced with systematic increments in order to evaluate its impact on hot carrier phenomena. At the same time, such an analytical approach was also attempted with the LDD structures in order to systematically vary the length of the lightly doped region. However, in both cases a deviation from the standard fabrication processes were required, which sometimes posed enormous difficulties to the production of the desired structures.

In Chapter 2 an introduction to the theory of hot carrier transport in semiconductors is presented, with emphasis on the methods used for calculating hot carrier generation and transport. Besides introducing the reader to the concept of hot carriers, this chapter also gives an idea of the importance of the electric field in the generation of hot

carriers. The discussion is limited to the transport in monocrystalline silicon, with no attempts to include the injection into SiO_2.

The IGFET is introduced in detail in Chapter 3. The Si/SiO_2 system is discussed, as well as the mechanisms and models for carrier injection. The role of the different biasing conditions in the hot carrier generation process, which in the past gave rise to controversy as to the preponderance of each type of carrier, is introduced with the aid of computer simulations. Finally, the two main diagnostic techniques used in this work, the charge pumping and and the gated diode methods, are reviewed.

The fabrication of offset-gate transistors and of systematically incremented LDDs needed special processes. They are presented in Chapter 4. An initial attempt was made to define LDDs by photolithographic techniques, which required further calibration. Due to the difficulties found in determining the exact LDD lengths, the inconsistencies found in some offset-gate arrays and in order to increase the range of transistor sizes, a second process was initiated, which used angled implants for producing the LDDs. Chapter 4 is devoted to the special characteristics of those two processes, as well as the various problems encountered in each fabrication step.

Both the gate offset and LDD increments had to be calibrated after fabrication. In some cases, like the photolithographically-defined LDD, no consistent results were obtained, even though as many as six different techniques, all with various biasing conditions, were used. This calibration is introduced in the first part of Chapter 5. It is followed by an account of the implementation of the different measurement methods. Finally, the measured results are presented with a brief discussion of their meaning.

For a better interpretation of the results gathered in Chapter 5, it is important to verify the significance of the gate offset in the degradation processes. The computer simulations used to explain the effect of different biasing conditions on degradation mechanisms are used again in Chapter 6 to examine this significance. Besides its significance in the electric field distribution and the density of available carriers, the presence of the gate offset is also discussed from the point of view of its impact in the analysis of some of the measurement results. With this background, the results shown in Chapter 5 are discussed and models for the degradation process and for the generation of interface

states are proposed.

Finally, Chapter 7 presents the conclusions and suggestions for future work.

References

1. Dennard, R. H., Gaensslen, F. H., Yu, H., Rideout, V. L., Bassous, E., and LeBlanc, A. R., "Design of Ion-Implanted MOSFETs With Very Small Physical Dimensions," *IEEE J. Solid-State Circuits*, vol. SC-9, pp. 256-268, October 1974.

2. Baccarani, G., Wordeman, M. R., and Dennard, R. H., "Generalized Scaling Theory and Its Application to a 0.25 Micrometer MOSFET Design," *IEEE Trans. Electron Devices*, vol. ED-31, pp. 452-462, April 1984. 007

3. Sangiorgi, E., Hofstatter, E. A., Smith, R. K., Bechtold, P. F., and Fichtner, W., "Scaling Issues Related to High Field Phenomena in Submicrometer MOSFETs," *IEEE Electron Device Lett.*, vol. EDL-7, pp. 115-118, February 1986. 020

4. Takeda, E., Jones, G. A. C., and Ahmed, H., "Constraints on the Application of 0.5 Micron MOSFETs to ULSI Systems," *IEEE J. Solid-State Circuits*, vol. SC-20, pp. 242-247, February 1985. 013

5. Baglee, D. A., Duvvury, C., Smayling, M. C., and Duane, M. P., "Lightly Doped Drain Transistors for Advanced VLSI Circuits," *IEEE Trans. Electron Devices*, vol. ED-32, pp. 896-902, May 1985.

6. Ogura, S., Tsang, P. J., Walker, W. W., Critchlow, D. L., and Shepard, J. F., "Design and Characteristics of the Lightly Doped Drain-Source (LDD) Insulated Gate Field-Effect Transistor," *IEEE J. Solid-State Circuits*, vol. SC-15, pp. 424-432, August 1980.

7. Tsang, P. J., Ogura, S., Walker, W. W., Shepard, J. F., and Critchlow, D. L., "Fabrication of High Performance LDDFETs with Oxide Sidewall-Spacer Technology," *IEEE Trans. Electron Devices*, vol. ED-29, pp. 590-596, April 1982.

8. Liou, T. I., Teng, C. S., and Merrill, R. B., "Hot-electron-induced degradation of conventional, minimum overlap, LDD and DDD n-channel MOSFETs," *IEEE Circ. Dev. Mag.*, pp. 9-15, March 1988.

9. Wei, C. Y., Pimbley, J. M., and Cohen, Y. N., "Buried and graded/buried LDD structures for improved hot-electron reliability," *IEEE Electron Device Lett.*, vol. EDL-7, pp. 380-382, June 1986.

Chapter 2 - Hot Carriers in Silicon

2.1. Introduction

Before examining the presence of hot carrier effects in IGFETs, it is important to know how hot carriers are originated and how they are affected by their environment. The very concept of a carrier being hot can be initially difficult to understand, considering that temperature is defined for atomic lattices rather than particles.

The transport of hot carriers in semiconducting silicon is one of the main issues in the study of hot carrier effects. In this Chapter, some of the main methods for the quantitative analysis of hot carrier transport are also discussed.

2.2. Macroscopic approach

The usual way of solving problems of device physics is to start by describing the device in terms of its electric field distribution. This field is created by all the external biases applied to the device plus that due to electric charges, which arise from the active impurity species and the free carriers that are always present at any temperature above absolute zero. Any change in electric potential should be proportional to the net concentration of those carriers. This can be written as

$$\frac{\partial^2 \phi}{\partial x^2} + \frac{\partial^2 \phi}{\partial y^2} + \frac{\partial^2 \phi}{\partial z^2} = -\frac{q}{\epsilon_S}(p - n + N_D - N_A) + \nabla . E_{ext} \qquad (2.1)$$

where q is the elementary charge, ϵ_S is the permitivitty in silicon, p and n are the hole and electron densities and N_D and N_A are the donor and acceptor concentrations. The carrier flow in this material has two origins: the flow caused by a gradient in carrier concentration (a diffusion mechanism) and the flow caused by the presence of an applied force. Therefore

$$J = (n \mu_n - p \mu_p) qF + q (D_n \nabla n - D_p \nabla p) \qquad (2.2)$$

where μ is the carrier mobility (which describes the carrier velocity dependence upon the applied electric field) and D is the diffusion constant. It is important to emphasise that carriers can only appear or disappear in the region under study through generation, recombination or injection to or from an adjacent region. This is called the *current continuity condition* and is written as

$$\nabla . J = G + R + \frac{\partial \rho}{\partial t} \tag{2.3}$$

The above equations describe what happens to a macroscopic fluid of current carriers when subjected to an electric field. They can accurately describe phenomena that do not affect the thermal equilibrium of the energy distribution of carriers because the carriers are not considered individually but as a continuous fluid. However, when it is necessary to analyse the behaviour of carriers which deviate from what would otherwise be regarded as average behaviour, those equations fail to predict many important effects.

2.3. Carrier statistics †

In a free electron gas, electrons wander randomly as a result of thermal agitation and exhibit a distribution of kinetic energies (and therefore of velocities) that follows a Maxwell-Boltzmann distribution

$$f(E) = \exp\left(-\frac{E}{kT}\right) \tag{2.4}$$

When the electron density is very high, the presence of quantum effects has to be taken into account. For example, the larger is the number of carriers, the larger the number of occupied energy states will tend to be, but as stated by Pauli, two electrons cannot occupy the same energy state simultaneously. Experimentally, it has been observed [1] that the Maxwell-Boltzmann distribution cannot account for certain properties of materials with a large number of conduction electrons, viz. the specific heat in metals. In such cases the Fermi-Dirac statistics must be used. Its distribution function is given by

$$f(E) = \left[1 + \exp\left(\frac{E - E_F}{kT}\right)\right]^{-1} \tag{2.5}$$

† From now on, it will be assumed that electrons predominate as current carriers; unless otherwise indicated, the application of the theory to holes is a simple matter of substitution with an appropriate change of sign.

where, for convenience, E_F is included to indicate the energy at which the probability of finding carriers is 0.5. When the difference [2] between E and E_F is of several kT, the distribution tail can be approximated by the Maxwell-Boltzmann equation (under such conditions the system is said to be nondegenerate).

In intrinsic semiconductors, the Fermi level is situated in the middle of the forbidden zone, half way from the valence to the conduction band. When the semiconductor is doped with donor impurities, there is an increase in the density of conduction electrons and the Fermi level shifts towards the conduction band, as depicted in Figure 2.1. The more it shifts towards the conduction band, the longer is the tail of the distribution situated above it. The nondegenerate approximation using the Maxwell-Boltzmann distribution is only valid for large values of $E - E_F$, so the system will eventually degenerate. At room temperature, $E - E_F$ has to be of less than about 0.1 eV to prevent us from using the nondegenerate approach; considering that

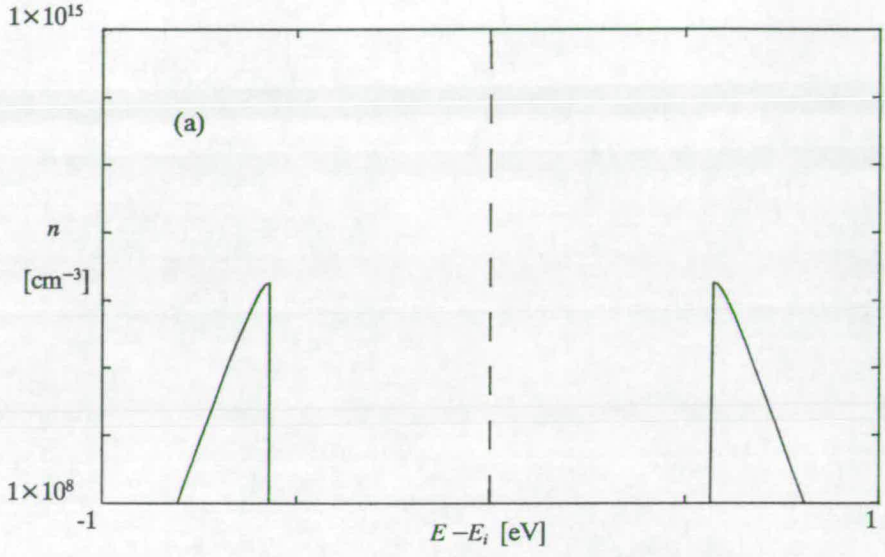
$$E_C - E_F = kT \ln \left(\frac{N_C}{N_D} \right) \tag{2.6}$$

where N_C is the effective density of states in the conduction band and N_D is the donor impurity concentration, and that at 300 K, $N_C \approx 2.8 \times 10^{19}$ cm^{-3}, the corresponding doping concentration has to be of above 10^{18} cm^{-3}. For example, in the present work, the IGFET channels have doping concentrations of the order of 10^{16} cm^{-3}, so we can easily consider the conduction electron distribution as Maxwellian under thermal equilibrium or low level injection.

The injection of carriers from silicon into the silicon dioxide gate insulator of an IGFET is only possible of the carriers have an energy at least equal to the potential barrier between those two materials ($\approx$3.2 eV). How many of those electrons are then available at room temperature? Considering an impurity concentration of 10^{16} cm^{-3}, the difference between E_H, the energy at which it is possible to surmount the barrier (measured with respect to the edge of the conduction band) and E_F will be

$$E_H - E_F = (E_C - E_F) + (E_H - E_C) = kT \ln \left(\frac{N_C}{N_D} \right) + 3.2 = 3.4 \text{ eV} \tag{2.7}$$

Using the Fermi-Dirac distribution to take advantage of the concept of E_F we have



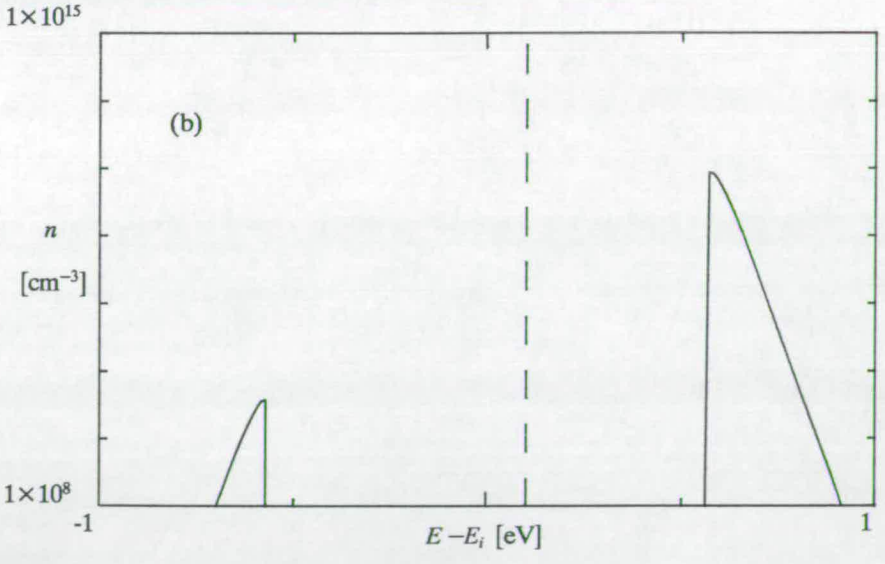


Figure 2.1: Density of electrons above and below the forbidden zone. (a) Intrinsic silicon; (b) n-doped silicon. In the latter, the Fermi level (marked with a dashed line) has shifted 0.1 eV towards the conductance band.

$$f(E)=\left[1+\exp\left(\frac{3.4}{2.59\times10^{-2}}\right)\right]^{-1}=7.6\times10^{-58} \tag{2.8}$$

The number of such electrons is also exceedingly small; rigorously, one would have to integrate

$$n=\int_{E_F}^{\infty} f(E)N(E)dE \tag{2.9}$$

in order to find it. However, even if we take the whole number of states in the conduction band (N_C) to estimate this value, it is evident that the final figure will be still negligible.

2.4. Hot electron generation

The electronic cloud constituted by conduction electrons has, as seen before, a random thermal motion. Any electric field applied to the system will create a systematic drift velocity superimposed on the thermal velocities. The drift velocity is much smaller than the thermal; in fact, the average speed in a Maxwellian distribution is given by

$$<s>=\sqrt{\frac{3kT}{m}} \tag{2.10}$$

which means that at 300 K this will be in the order of 10^5 m/s. For an applied field of 10^4 V/m, which is low enough not to affect the thermal equilibrium, the drift velocity will be

$$v_d=\mu E=1.5\times10^{-1}\ 10^4=1.5\times10^3\ \text{m/s} \tag{2.11}$$

The crystal lattice where the electronic flow takes place is vibrating as a consequence of the temperature. Silicon has a diamond structure, which can be regarded as the result of two interpenetrating face-centred cubes [3] (Figure 2.2). It can therefore be seen as a face-centred cube with a diatomic basis. Such a structure has two main modes of vibration. The first is characterised by the relative vibration amongst different bases in a face-centred cube (Figure 2.3.a). The energy associated with such vibrations, like any other lattice vibration, is quantised. A quantum of lattice vibration is called *phonon;* in this particular case of inter-basis oscillation, the phonons are called *acoustic phonons*.

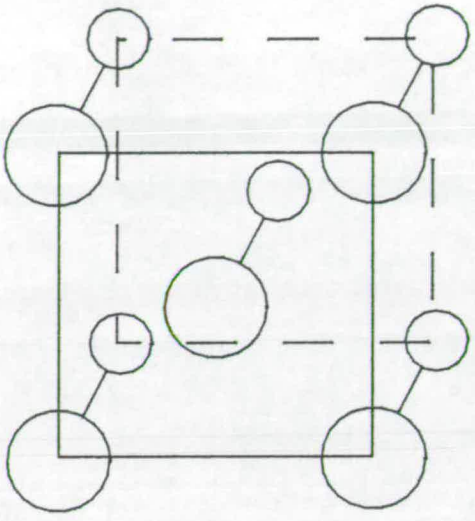


Figure 2.2: One of the faces of the face-centred cube silicon structure. Each of the five atoms in this face (marked with a solid square) has another atom (connected by a solid line) with which it constitutes a diatomic basis; consequently the companion atoms also constitute a parallel face (marked with a dashed square). The same situation extrapolates to the other faces of the face-centred cube and the two resulting cubes can be seen as inter-penetrating.

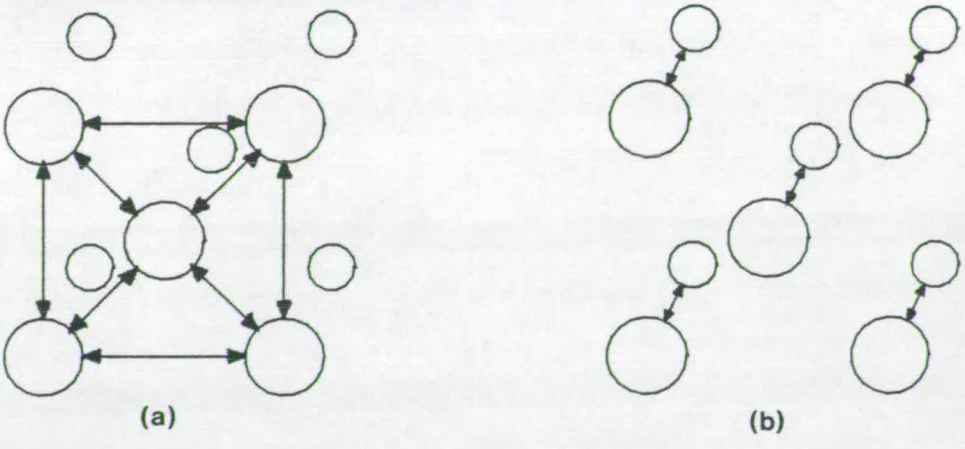


Figure 2.3: Vibration modes of the silicon structure. (a) Acoustic mode (vibration between atoms of a same face-centred cube); (b) Optical mode (vibration within bases).

The second mode of vibration originates from the oscillation within the basis (Figure 2.3.b): it is an intra-basis mode as opposed to the inter-basis oscillation of the acoustic mode. The quantisation of this second mode gives *optical phonons*. Acoustic phonons exist in a large number and are characterised by having a long wavelength, which implies in a low energy. Optical phonons, on the other hand, are much more scarce and have a shorter wavelength (and therefore higher energies).

Electrons gain energy when an electric field is applied. This energy is usually lost through collisions with the vibrating lattice. This may sound contradictory when we consider the fact that matter waves should propagate freely in a periodic structure,[3] but the lattice oscillations intermittently break this regularity. Such collisions can be seen as the interactions between electrons and the compressions and rarefactions in the lattice potential.[1] The important condition for these collisions to take place is that the phonons should have high enough energies so that their wavelength ($E = h\,\omega_{ph}$) is at least of the same order of magnitude as the electron wavelength.

When moderate fields are applied, electrons will tend to lose any energy gained in collisions with acoustic phonons. Although the large number of acoustic phonons will make such collisions very likely to occur, their low energies will signify a relatively low energy loss per collision. The system will still be under equilibrium, but not at the same lattice temperature T. An electron temperature T_e should therefore be used to characterise such a distribution which is still Maxwellian. The dependence of the drift velocity upon the electric field (Figure 2.4) starts to deviate from a straight line (whose slope is the mobility) according to [4]

$$v_d = \mu E \sqrt{\frac{T}{T_e}} \tag{2.12}$$

The critical electric field E_c at which this deviation from linearity takes place is given by [4]

$$E_c = \frac{32 c_s}{3\pi\mu} \tag{2.13}$$

where c_s is the speed of sound in silicon (= 9150 m/s). E_c has therefore a value of 8×10^4 V/m.

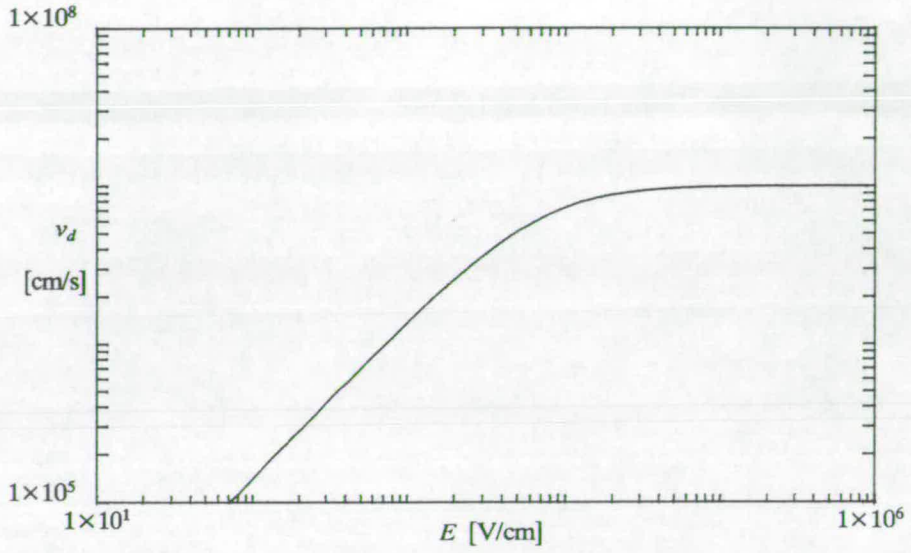


Figure 2.4: Variation of the drift velocity with the applied electric field.

If the field is further increased, electrons will gain even more energy and it will become increasingly difficult for them to lose this energy in collisions with acoustic phonons. This increasing energy excess will eventually make the electrons reach the optical phonon energy (0.063 eV for silicon). This will be equivalent to an electron temperature of about 730 K. Entering this figure in the Fermi-Dirac distribution as done before for the calculation of the probability of finding injection electrons at room temperature, we have that

$$f(E)=[1+\exp(\frac{3.4}{kT_e})]=3.6\times10^{-24} \tag{2.14}$$

Therefore there was an increase of the order of 10^{33} times in this probability. These interactions with optical phonons will almost saturate the mobility curve; however, due to the limited number of optical phonons available for collision, some electrons will have the possibility of increasing their energies above the optical phonon energy,[1] giving rise to a further small increase in drift velocity. Electrons that reach this region of almost saturation in drift velocity are called *hot electrons*.

If the electric field is increased even more, electrons can generate electron-hole pairs through impact ionisation. This is an avalanche mechanism, i.e. carriers generated by it will in turn generate new pairs and so on. This will abruptly increase the electron velocity with almost no change in the applied field. Based on the premise that impact ionisation is created mainly by non-equilibrium electrons,[5] Shockley [6] developed the *lucky-electron model,* in which he states that impact ionisation is produced by electrons which were lucky not to lose their energies through collisions. His study was based in the use of three constants, which are empirically adjusted to fit experimental data: E_i, the energy necessary for impact ionisation, L_R, the mean free path between electron - optical phonon collisions (also called Raman scatterings) and r ($\equiv L_i/L_R$, where L_i is the mean free path for electrons whose energies are greater than E_i). According to Shockley, the number of collisions required for an electron to lose its energy is

$$C=\frac{E-E_i}{E_R} \tag{2.15}$$

and in the process of losing this energy it travels a distance equal to CL_R. Therefore, the probability of not causing impact ionisation whilst losing energy should be

$$\bar{P} = \exp\left(-\frac{CL_R}{L_i}\right) = \exp\left(-\frac{C}{r}\right) = \exp\left[-\frac{(E - E_i)}{rE_R}\right] \tag{2.16}$$

Thus the ionisation probability is

$$P = 1 - \bar{P} = 1 - \exp\left[-\frac{(E - E_i)}{rE_R}\right] \tag{2.17}$$

Shockley points out that the oversimplification of this model eventually results in the loss of the physical meaning of the adjustable constants which had been initially assigned.

As we can see, this lucky electron approach consists of a very simple probability model. In the case if IGFETs, where not only the generation of hot carriers but also their injection into SiO_2 needs to be examined, the model can easily incorporate the probabilities of other participating mechanisms. This will be further discussed in Chapter 3.

2.5. The Boltzmann transport equation

As has been pointed out in the previous sections, the electrons of interest for this study will usually be localised in the extreme tail of the Maxwell-Boltzmann distribution. This means that we should observe what happens to this distribution as the electronic ensemble is transported across the device channel. More precisely, we are looking for a solution for the distribution $f(x, p, t)$ as a function of energy when it varies in time. This should show the variation in the hot electron population as transport takes place. The equation which describes such transport is the *Boltzmann transport equation*.

The variation in time of the function $f(x, p, t)$ is given, in common space, by

$$\left.\frac{df}{dt}\right|_{C.S.} = \frac{\partial f}{\partial x}\frac{dx}{dt} = \frac{\partial f}{\partial x}v = \frac{\partial f}{\partial x}\frac{p}{m^*} \tag{2.18}$$

and similarly, in momentum space, by

$$\left.\frac{df}{dt}\right|_{M.S.} = \frac{\partial f}{\partial p}\frac{dp}{dt} = \frac{\partial f}{\partial p}m\frac{dv}{dt} = \frac{\partial f}{\partial p}F \tag{2.19}$$

Combining the equations above, we have the generic form of the rate of change in time of $f(x, p, t)$

$$\frac{df}{dt} = \frac{\partial f}{\partial x}\frac{p}{m^*} + \frac{\partial f}{\partial p}F + \frac{\partial f}{\partial t} \tag{2.20}$$

The previous section describes some of the mechanisms through which electrons lose their energy. Such events cannot be incorporated in the applied electric field [7] and therefore will need an extra term in Equation 2.20. There are different ways to take collisions into account; in fact, this can be one of the major complicating factors that will make solution for the Boltzmann transport equation extremely difficult and sometimes impossible. The most common way to deal with collisions is by using the relaxation time approximation.[7, 8] This approximation is based on the assumption that any change in the equilibrium distribution f_0 towards a new distribution f will be restored within the relaxation time. For the purpose of low field analysis, one can make use of the mean relaxation time τ and write the collision term as

$$\frac{df}{dt} = \frac{f - f_0}{\tau} \tag{2.21}$$

In a more general situation, however, it must be considered that there is a nonuniform probability distribution of collisions in momentum space, i.e. the mean relaxation time is not a constant but a function of p. The formula for this approach then becomes [9]

$$\frac{df}{dt} = \int f(p')W(p' \to p)dp' - \int f(p)W(p \to p')dp \tag{2.22}$$

where the relaxation time appears in the form of two distinct scattering times

$$\tau = \frac{1}{\int W(p \to p')dp} \quad \text{and} \quad \tau' = \frac{1}{\int W(p' \to p)dp'} \tag{2.23}$$

The collision expression therefore describes distinctively the scattering out and into state p.[10]

The final form of the Boltzmann transport equation is

$$\frac{\partial f}{\partial x}\frac{p}{m^*} + \frac{\partial f}{\partial p}F + \frac{\partial f}{\partial t} = \int f(p')W(p' \to p)dp' - \int f(p)W(p \to p')dp \tag{2.24}$$

which is an integro-differential equation. Such equations are normally impossible to solve analytically,[11] therefore drastic approximations are necessary. In any case, it is important to remember that, even in the most complete form, there are severe limitations to its use as an analytical tool for hot electron phenomena. Amongst such limitations, it neglects several different aspects of the scattering processes (such as the forces

which accelerate electrons during scattering) and the concomitant presence of various scattering mechanisms with different probabilities. It also assumes that temperature is a constant parameter, which is particularly detrimental when the concept of electron temperature (as opposed to lattice temperature) comes into effect. Additionally, the relaxation time approach is only applicable when the energy change is small enough compared to the carrier energy. As we have already seen, this is not true for collisions with optical phonons.

2.6. The Monte Carlo technique

The task of looking at nonlinear transport (i.e. hot electron transport) using current equations can be compared to the use of a magnifying glass to observe a microscopic phenomenon. An overall view is possible but no finer details can be studied. The solution of the Boltzmann equation would then be equivalent to using a microscope without enough magnification power to look at the details which we are interested in. The suitable "super-microscope"[12] is, as we will see, the Monte Carlo technique.

The Monte Carlo method consists of calculating the movement of individual carriers (or groups of carriers) in a device. The transport between interactions with phonons, the interactions themselves, the resulting angles, all the stochastic mechanisms are computed using a pseudo-random number generation. This presents one of the primary difficulties of the method: one has to select times, processes and angles out of nonuniform distributions by choosing uniformly distributed random numbers. Furthermore, the heuristics involved has to be accommodated in a reasonably short computing time, so that more particles can be incorporated to the solution to allow a more reallistic approach.

The first step is to compute the electric field mesh using Poisson's equation, taking into consideration the device biases, active impurities, fixed and mobile charges, etc. There are several methods to do so and the most commonly used for this particular application is known as POT4,[13] which is based on Fourier analysis together with cyclic reduction. As the computation of carrier transport evolves towards its final solution, there will be a change in field distribution due to the redistribution of the carriers

themselves. A periodic re-calculation of the electric field can then be performed; this leads to the self-consistent Monte Carlo technique.[12]

Having calculated the field distribution, we then select the particles one by one in order to compute their transport through the field mesh. The flight time, or the time during which the carrier will travel without suffering collisions, is determined using the probability of a collision to occur between t and dt, which is given by

$$P(t)dt = \Gamma \exp\left[-\int_0^t \Gamma dt'\right] dt \tag{2.25}$$

where Γ is the sum of the scattering rates λ_i of all the possible interactions for a given wave vector (for λ_i is a function of k). It happens, though, that Γ is not even linear, which makes the problem computationally difficult. The practical solution encountered is to introduce a ficticious interaction, called *self-scattering*, so that

$$\Gamma_0 = \Gamma + \lambda_{self-scattering} = \text{const.} \tag{2.26}$$

Figure 2.5 shows an example of the use of the self-scattering mechanism. A self-scattering interaction should, of course, be understood as an interaction in which there is no change in carrier energy or momentum. Considering the new Γ_0 we now have

$$P(t) = \Gamma_0 \exp(-\Gamma_0 t) \tag{2.27}$$

Solving the above equation for t gives

$$t = \frac{1}{\Gamma_0} \ln\left[\frac{\Gamma_0}{P(t)}\right] \tag{2.28}$$

Since $P(t)$ is a probability whose value lies between 0 and Γ_0, we have that

$$0 \leq \frac{\Gamma_0}{P(t)} \leq 1 \tag{2.29}$$

Therefore we can use a random number r, between 0 and 1, to choose the time of free flight. The new wave vector and position after the flight are calculated from

$$k'_x = k_x + \frac{eE_x t}{\hbar} \tag{2.30}$$

$$k'_y = k_y + \frac{eE_y t}{\hbar} \tag{2.31}$$

$$x' = x + \frac{\hbar}{2m}(k'_x + k_x)t \tag{2.32}$$

$$y' = y + \frac{\hbar}{2m}(k'_y + k_y)t \tag{2.33}$$

After a free flight, a collision is supposed to take place. The choice of the scattering mechanism is made using a second random number, r_2. The choice of a process p means that

$$\sum_{i=0}^{p-1} \lambda_i(k) \leq r_2 < \sum_{i=0}^{p} \lambda_i(k) \tag{2.34}$$

For the sake of saving computer time, the values of $\lambda_i(k)$ are usually tabulated in the memory space before each run. However, there is still another inconvenience to circumvent: the fact that self-scattering has a risk of being chosen too often. In Figure 2.5, we see that the distance between any two probabilities of real processes is sometimes much less than the distance between the probability of self scattering and the uppermost probability of a real process, particularly for low values of k. This can be avoided by optimising the value of Γ_0.[13, 14] It is chosen so that

$$\Gamma_0 = \sum_{i=0}^{p} \lambda_i[k(0)] + \Delta\Gamma \tag{2.35}$$

where $k(0)$ is the wave vector at the start of a flight. $\Delta\Gamma$ is initially set to zero, so that Γ_0 is initially made equal to the sum of all scattering rates up to the greatest, considering the initial wave vector. This value of Γ_0 is used for computing the flight time as well as the new wave vector and position (as seen before). If the sum of the scattering rates taken with respect to the new wave vector is less than Γ_0, then $\Delta\Gamma$ is kept the same and a scattering process is chosen; otherwise we increase $\Delta\Gamma$ and re-calculate the flight time. This procedure is represented within the dashed rectangle of Figure 2.6, which shows the flowchart for the entire Monte Carlo process.

Finally, the results of the scattering process have to be evaluated. Except for self-scattering, there is a change in direction of flight as well as in the wave vector. The spread of possible directions depends on the scattering mechanism involved. The interaction between electrons and acoustic phonons, for example, is totally randomising; therefore, it is sufficient to choose a new random number r_3 and make the resulting angle

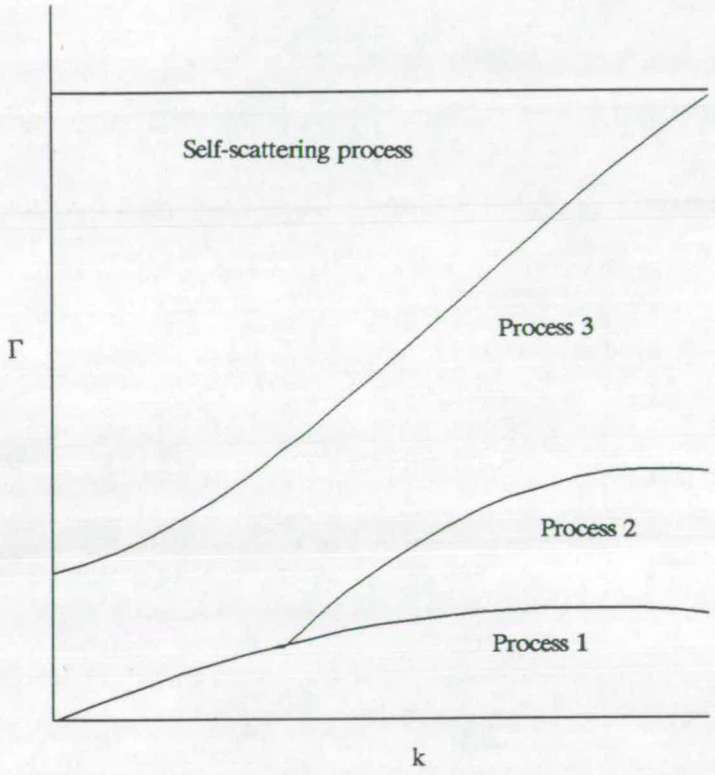


Figure 2.5: Example of the use of a self-scattering process. Three scattering processes are represented in this scattering rate versus wave vector graph. Because the total scattering rate varies with k, a dummy process is added so that the total rate becomes constant. This procedure simplifies the random choice of a flight time.

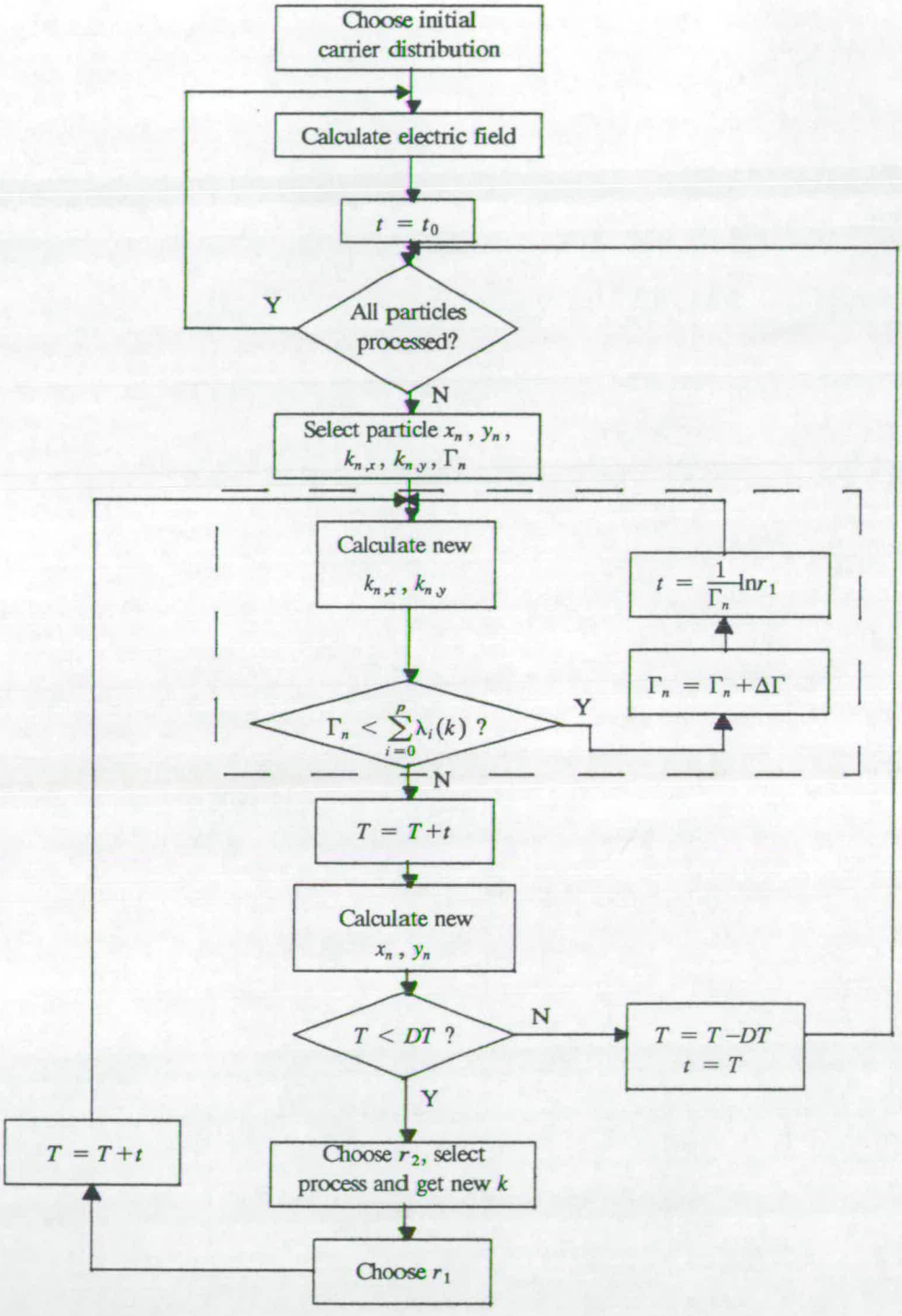


Figure 2.6: Flowchart of a typical Monte Carlo particle calculation. DT is the time interval required for simulating one scattering and flight time for the whole set of particles. The portion enclosed in the dashed box adjusts the choice of the self-scattering rate.

$\theta = 2\pi r_3$. On the other hand, if the mechanism being considered is optical phonon scattering, it is necessary to compute the angle as a function of the previous k.

The Monte Carlo technique constitutes an excellent tool for a microscopic analysis of transport phenomena, providing instantaneous pictures of carrier distributions and related quantities. The technique can account for as many interaction mechanisms as we wish and, by choosing a suitable number of carriers and adequate time intervals, can be sufficiently chaotic to reproduce quite well the complexity of the transport processes. However, it is important to bear in mind that the consistency of results depends primarily on the accuracy and order of approximation of the informations being input to the problem. Sometimes the limits are imposed by our own understanding of the mechanisms involved. For example, the injection of electrons from silicon into the silicon dioxide insulator of IGFETs can be included in a Monte Carlo simulation only as precisely as our knowledge of the phenomenon permits. In such cases, and particularly when departures from the idealised picture of flawless crystals and perfect materials occur, theory ceases to be the absolute supplier of data. Whatever the solution method is, a sound set of experimental results is absolutely necessary.

References

1. Blakemore, J. S., *Solid State Physics,* Cambridge University Press, Cambridge, 1985.
2. Seeger, K., *Semiconductor Physics,* Springer-Verlag, Berlin , 1982.
3. Kittel, C., *Solid State Physics,* John Wiley & Sons, New York, 1976.
4. Moll, J. L., *Physics of Semiconductors,* McGraw-Hill, New York, 1964.
5. Ridley, B. K., "Lucky-drift mechanism for impact ionisation in semiconductors," *J. Physics C: Solid State Phys.*, vol. 16, pp. 3373-3388, 1983.
6. Shockley, W., "Problems related to p-n junctions in silicon," *Solid State Electron.*, vol. 2, pp. 35-67, 1961.
7. Harrison, W. A., *Solid State Theory,* McGraw-Hill, Tokyo, 1970.

8. Carroll, J. E., *Rate Equations in Semiconductor Electronics*, Cambridge University Press, Cambridge, 1985.

9. Price, P. J., "The theory of hot electrons," *IBM J. Res. Develop.*, vol. 14, pp. 12-24, January 1970.

10. Law, H. C. and Kao, K. C., "Analytical solutions of the Boltzmann transport equation," *Phys. Rev. B*, vol. 4, pp. 2524-2534, 1971.

11. Cercignani, C., *Theory and Application of the Boltzmann Equation*, Scottish Academic Press, Edinburgh, 1975.

12. Moglestue, C., "A self-consistent Monte Carlo particle model to analyze semiconductor microcomponents of any geometry," *IEEE Trans. Computer-Aided Design*, vol. CAD-5, pp. 326-345, April 1986.

13. Hockney, R. W. and Eastwood, J. W., *Computer Simulation Using Particles*, McGraw-Hill, New York, 1981.

14. Kato, K., "Hot-carrier simulation for MOSFETs using a high speed Monte Carlo method," *IEEE Trans. Electron Devices*, vol. ED-35, pp. 1344-1350, August 1988.

Chapter 3 - Hot carrier investigation in IGFET devices

3.1. Introduction

The generation of hot carriers and their subsequent injection into the gate insulator are the two key elements in the analysis of hot carrier phenomena in IGFET devices. Although the theory of hot carrier generation and transport has led to reasonably accurate solutions (as discussed in Chapter 2), they invariably have to rely on simplifications which sometimes differ too much from the real case. In IGFETs the situation is even more serious because of the two-dimensional character of hot carrier generation and injection. Furthermore, once the carriers are injected, they are subject to the complexity of the Si/SiO_2 interface and of the not-so-perfect SiO_2 structure, where they can not only become trapped but also generate new traps. It is therefore important to find answers for the following questions:

a) Which conditions can lead to hot carrier generation in IGFETs?

b) What is the energetic and spatial distribution of hot carriers?

c) What are the possible injection mechanisms?

d) How does injection qualitatively and quantitatively affect the insulator (interface state generation, charge trapping, etc.)?

e) How is device reliability affected by hot carriers?

f) How can we avoid device aging by hot carriers?

In attempting to answer some of these questions, many different techniques were developed; some of the most relevant are discussed in this Chapter.

However, before discussing the methods for hot carrier analysis, it is important to look at the physical phenomena involved in injection and to relate IGFET operating characteristics to hot carrier generation.

3.2. Hot carrier injection into SiO_2

As we have discussed in Chapter 2, we call hot carriers those which, after gaining energy, can be situated several *kT* above the Fermi level and are no longer in thermal equilibrium with the lattice. Such carriers become a problem for IGFET operation when they reach energies high enough to allow them to surmount the Si/SiO_2 barrier and get injected into the SiO_2. Figure 3.1 shows the band diagram of the polysilicon-gate/SiO_2/Si system. One can immediately extract from this diagram some useful facts which have a direct bearing upon hot carrier injection. Under such equilibrium conditions, there are three different ways by which an electron could be injected from Si into SiO_2: to go through the barrier (path "1" in Figure 3.1), to gain some energy and go through the thinner triangular region at the top of the barrier (path "2" in Figure 3.1) or to gain enough energy to surmount the barrier (path "3" in Figure 3.1). All three mechanisms can occur, depending on how much energy the electrons can gain and how thin the barrier is (i.e. how thin the SiO_2 layer is). The first mechanism is called *tunneling*, the second *Fowler-Nordheim tunneling* and the third is the real *hot carrier injection*.

The diagram also shows that both hot electron and hot hole injection can take place. If a hole with high enough energy appears in the valence band near the Si/SiO_2 interface then it can be transferred into the valence band of SiO_2. However, as the diagram shows, this hot hole injection is less likely to occur than hot electron injection due to the higher energy required (4.7 eV for holes as opposed to 3.2 eV for electrons).

In an ideal insulator, carriers will drift towards the gate conductor once they surmount (or cross through) the barrier. However, in real insulators there are many possible obstacles, or defects, to stop them. The injected carriers themselves can also introduce some of these defects. Thermally grown oxide, which has a vitreous structure (i.e., the structure is fairly regular within a few atomic separations but irregular over long lengths [1]) has a variety of distorted Si-O bonds. Some are excessively elongated, others are strained, rotated or even broken. There are unbound Si and O atoms, as well as foreign ions such as Na^+ and hydrogenated compounds. At the interface between SiO_2 and Si there are lattice discontinuities which lead to dangling bonds.

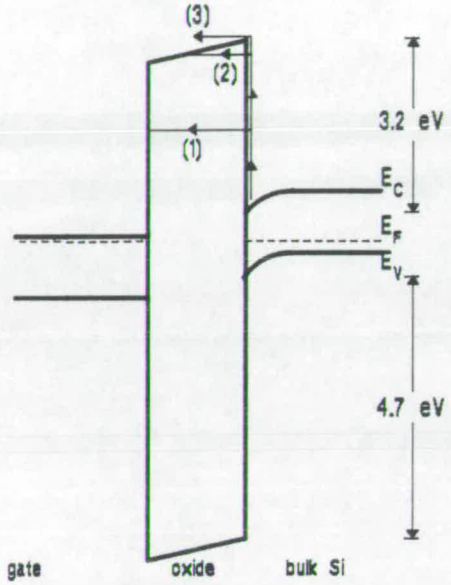


Figure 3.1: Band diagram of the polysilicon-gate/SiO_2/silicon system under thermal equilibrium.

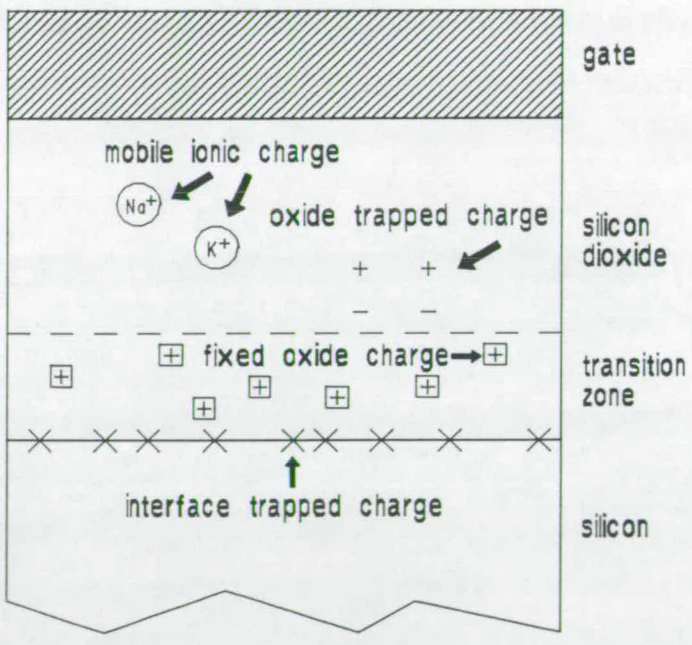


Figure 3.2: Classification of the SiO_2 charges (after [1]).

Depending on the nature of the defect, they can have an electric charge (positive or negative) or be neutral. A dangling silicon bond, for example, will behave like a positive charge. The likelihood of a carrier being trapped by a defect therefore depends not only on the defect size but also on its charge. The concept of *capture cross section* conveys both concepts. It gives a measure of the "area of action" of a defect. For instance, the capture cross section for an electron by a positively charged trap is greater than by a neutral trap.

According to their origin and characteristics, defect charges are classified [2] as *mobile ionic charges, oxide trapped charges, fixed oxide charges* and *interface trapped charges* (also known as interface states). They are depicted in Figure 3.2. Mobile ionic charges are caused mainly by foreign impurities and are present over the entire volume of the thermally grown SiO_2. Fixed oxide charges are usually process-related structural imperfections and are located in the transition region (approximately 25 Å thick) between the interface and the bulk of SiO_2. Interface trapped charges, as the name implies, are located at the interface itself and can be process-related or caused by the breakage of bonds by different agents, such as injected carriers. Oxide trapped charges are situated deeper than fixed oxide charges and consist of carriers which have been injected and trapped because of structural defects. Defects which are situated at the interface between silicon and SiO_2 are particularly important because they can easily exchange carriers with the bulk silicon. Depending on how fast this carrier transfer occurs, they are classified as *fast states* or *slow states*. Fast states are normally charged and discharged within a few milliseconds or less, whilst in slow states it can take from seconds to days or months.[3]

The presence of defects in the SiO_2 will affect the operating characteristics of IGFETs by changing the surface potential due to the trapped charges. This will result in shifts in the flatband and threshold voltages. Additionally, the presence of interface traps can introduce additional energy levels within the silicon bandgap at the Si/SiO_2 interface. Like the impurity species used for doping silicon, these defects can behave like donors or acceptors due to the ease in transferring carriers to and from the bulk silicon, as discussed in the previous paragraph. Acceptor-type defects are situated between the valence band and midgap. They are neutral when empty or negatively charged when an electron

is trapped. Donor-type defects are situated between midgap and the conduction band and are neutral when filled by an electron and positively charged when empty. The presence of charges associated with acceptor and donor-type interface traps will therefore depend on the surface potential: every time the trap energy level is lowered below the Fermi level the trap will be filled and when the trap energy level is raised again the trap will be emptied. Such charges cause shifts in the flatband and threshold voltages which will vary with the applied gate bias; the charges will behave like a shield to the gate bias.[4] For example, let us consider the situation of an n-channel transistor with a concentration of acceptor-type interface states in the SiO_2 insulator. As the gate bias is increased in order to drive the transistor into depletion, the interface states are lowered below the Fermi level and become negatively trapped. Because negative charges are already being provided by the interface states, there will be less band bending and the device will be more difficult to deplete, leading to a higher threshold voltage.

Two different models are usually used for hot carrier injection: the *lucky electron* model, adapted from Shockley's empirical model for impact ionisation, and the *effective electron temperature* model. Shockley's model for impact ionisation, already discussed in Chapter 2, is based on the probability of an electron travelling a distance d without losing its energy through collisions. This probability can be written as $P = \exp(-d/\lambda)$, where λ is the electron mean free path. For the modelling of the injection mechanism, this probability (which represents the hot electron generation process) is multiplied by the probability of not suffering energy-losing collisions on its way before being injected. Although it is widely used (despite the lack of physical meaning of its parameters once they are fitted to experimental data), this model has one major disadvantage: it is a one-dimensional model and cannot accurately describe injection processes with a spatial distribution, such as those involved in IGFETs.[5] Nevertheless, the model has been used successfully to predict injection if it is assumed that isotropic scattering is taking place prior to injection and the maximum channel field is known.[6]

The effective electron temperature model is based on the assumption of a Maxwellian distribution of hot electron energies, from which it is possible to assume that there is an effective temperature T_e higher than the lattice temperature. Once this assumption is made, the problem of injection can be treated as simple thermionic emission, with the

emission probability given by [7]

$$P \propto \exp(\Delta\phi/kT_e) \tag{3.1}$$

where $\Delta\phi$ is the image force barrier lowering potential. The difficulty, however, is that T_e is difficult to compute. It depends on the gradient of the electric field rather than its local value. It is calculated as [8]

$$T_e(x) = \frac{2q}{5k}\int_0^\infty E_x(x-u)\exp(-3u/5\tau_e v_s)du \tag{3.2}$$

where E_x is the horizontal component of the electric field, τ_e is the energy relaxation time and v_s is the saturation velocity. The calculation of T_e becomes even more difficult under the field distributions which lead to hot carrier generation in IGFETs because of their high gradients.

3.3. IGFET biasing conditions for injection

In order to produce hot carriers in an IGFET device, the only necessary condition is that a high enough electric field be present. There are three different device biases in which this condition can be satisfied: a) $V_{DS}=0$, $V_{GS}>V_T$ and $V_{BS}<0$; b) $V_{DS}\approx V_{GS}$ (i.e. near saturation); and c) $V_{DS}>>V_{GS}$ (heavy saturation). The first case leads to what is known as *substrate hot carriers* and the other two to *channel hot carriers*. Channel hot carrier effects are the most common source of hot-carrier-related degradation of operating parameters in IGFETs (also referred to as *aging*), whilst substrate hot carriers can be avoided by a choice of a suitable substrate bias.

The bias condition for substrate hot carrier injection can be compared to a MOS capacitor, with the exception that it can be kept under non-equilibrium (i.e. deep depletion) conditions for as long as one wants because any minority charge build-up will be drained by the source/drain junctions (Figure 3.3). Carriers generated in the depletion region will be accelerated by the high electric field near the interface. Depending on the biasing voltages, the number of hot carriers can be considerably increased by avalanche multiplication.

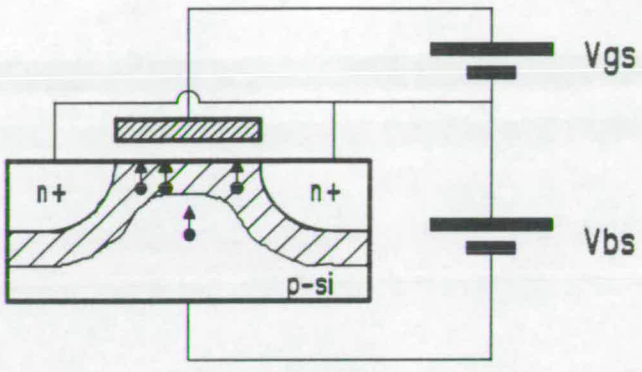


Figure 3.3: Substrate hot electron generation. Electrons present in the substrate (e.g. generated in the depletion region) are accelerated towards the Si/SiO_2 interface by the high electric field in the depletion region.

Although substrate hot carrier effects can be easily avoided, they have been extensively used in hot carrier research.[9] This is due to the one-dimensional character of the injection process, particularly if the device dimensions are made large compared to the width of the depletion layer. Additionally, the analysis of the hot carrier injection process can be greatly facilitated by the supply of a large number of such carriers instead of having to rely on the scarce availability of generated minority carriers. The possibility of driving the surface into the avalanche regime is one of the ways to circumvent it. Another method for increasing the density of carriers is to make devices with an underlying junction.[10] Minority carrier generation can also be stimulated using light.[11] It is important to emphasise, however, that the use of substrate hot carriers is for the sole purpose of studying injection mechanisms; in such cases the specific issue of hot carrier aging in IGFETs is not being addressed.

Whenever IGFETs are biased either in the linear or in the saturation region, there is a chance of channel hot electron generation. It only depends on how the electric field is distributed and how high it is. In this case, the horizontal electric field will be responsible for the hot carrier generation, in contrast with the predominantly vertical transport observed in the case of the substrate hot carriers. Because the field responsible for carrier heating has two different origins, one horizontal (due to the drain bias) and one vertical (due to the gate bias), at least part of the hot carrier transport and/or injection has a two-dimensional character, which makes it more difficult to analyse. Depending on whether the device is operated in the linear or saturated regions, this injection can assume different characteristics.

One of the most common bias conditions for IGFET transistors in VLSI circuits is to set $V_{GS} \approx V_{DS}$. As can be seen in the simulation of Figure 3.4 (produced with the CANDE device simulator), the potential stays more or less constant over most part of the channel and increases gradually as the drain junction is approached. This can be further confirmed by Figure 3.5, which shows the values of both the horizontal and vertical components of the electric field across the channel. The vertical field is almost constant from source to drain and there is a zero horizontal field apart from a sudden increase very near the drain junction. This is the situation we find in a long channel transistor: a purely vertical electric field over most of the channel length except for the

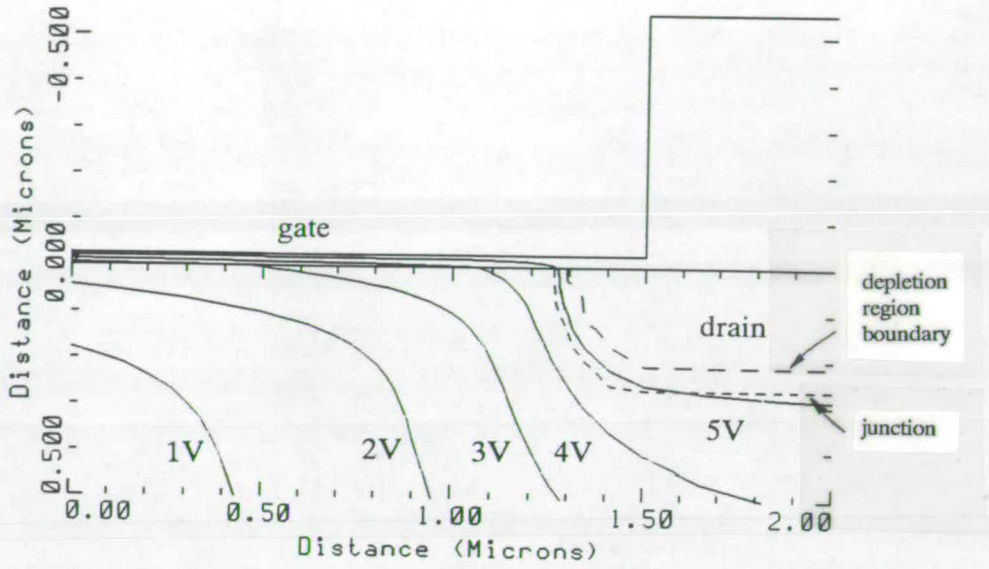


Figure 3.4: Simulation of the potential distribution (solid lines) near the drain junction of a 5μm n-channel transistor. $V_{GS} = V_{DS} = 5V$. The increment between adjacent lines is 1V.

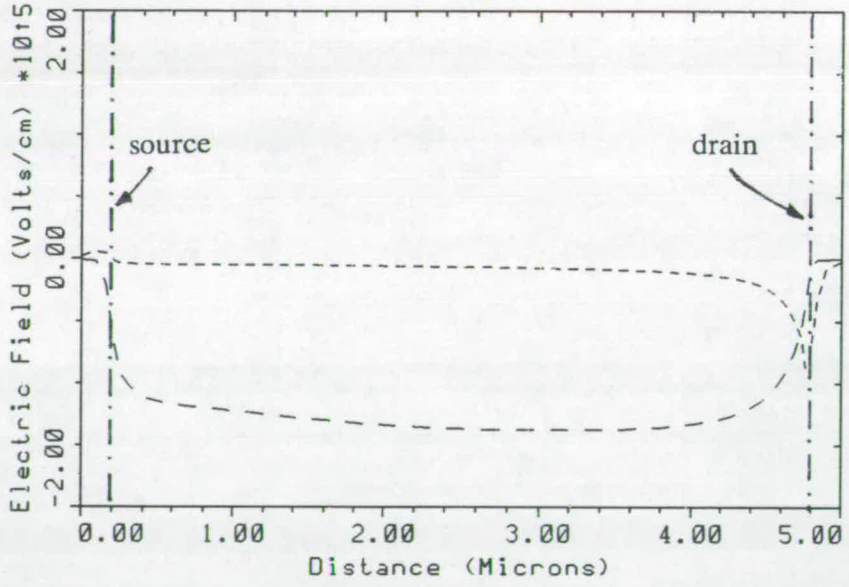


Figure 3.5: Horizontal (short-dashed line) and vertical (long-dashed line) components of the electric field across the channel of the transistor of Figure 3.4.

region immediately adjacent to the drain. Under such circumstances the field is quite uniform and the small perturbation introduced by the drain bias in the horizontal field near the drain is not likely to cause any carrier heating. This situation can change drastically as transistors are made smaller. Figure 3.6 shows the equipotential lines for the same bias on a shorter transistor (1 μm channel length as opposed to the 5 μm case discussed before). The lines near the Si/SiO_2 interface now appear much more perpendicular to the electron path than in the first case. The horizontal electric field now increases gradually in the middle part of the channel and it peaks at a much higher value at the drain, as we can see in Figure 3.7. In contrast to the long channel situation discussed above, the horizontal field is strongly affected by the drain bias and even the vertical field is distorted by it. The electric field distribution is essentially two-dimensional over the entire channel. As electrons cross the channel, they see a moderate vertical electric field which is not enough to heat them. As they approach the drain junction, the electric field suddenly increases over a short length; therefore they are likely to get heated and not to lose their energy through collisions over such a short path. A proportion of those electrons can therefore be injected towards the SiO_2 interface. This is eminently a hot electron injection process. Since no avalanche process is involved, the number of such carriers is small but the probability of injection is high (given that there are no repelling fields).

As the transistor is driven into saturation by reducing the gate voltage, pinch-off occurs and a depletion layer builds up between the pinch-off point and the drain. This channel length modulation by the gate voltage can be quantitatively defined (using the Reddi-Sah model [4]) as

$$\Delta L = \sqrt{\frac{2\epsilon_{Si}}{qN_A} V_{DS} - V_{DS,SAT}} \tag{3.3}$$

where the dependence of $V_{DS,SAT}$ on V_{GS} is given (using the Ihantola-Moll model [4]) by

$$V_{DS,SAT} = V_{GS} - V_{FB} + 2\phi_B + \frac{\epsilon_{Si} q N_A}{C_{ox}^2} \left(1 - \sqrt{1 + \frac{2(V_{GS} - V_{FB}) C_{ox}^2}{\epsilon_{Si} q N_A}} \right) \tag{3.4}$$

As a consequence of the shortening of the inverted channel, the flux of electrons in the region between the pinch-off point and the drain region is no longer confined to the

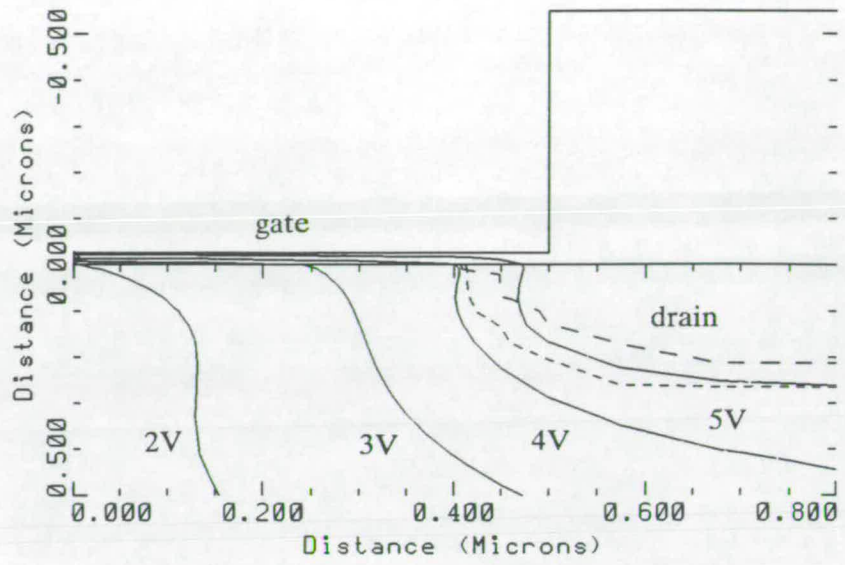


Figure 3.6: Simulation of the potential distribution (solid lines) near the drain junction of a 1μm n-channel transistor. $V_{GS} = V_{DS} = 5V$. The increment between adjacent lines is 1V.

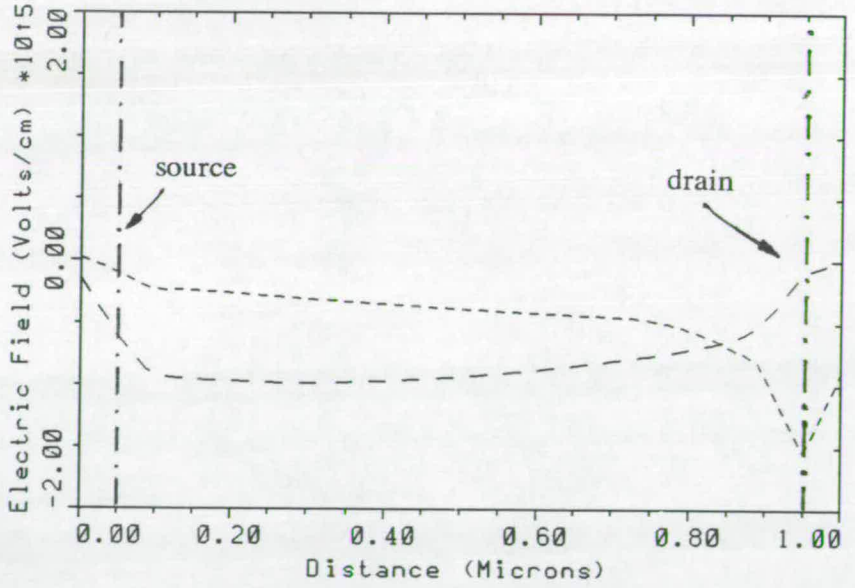


Figure 3.7: Horizontal (short-dashed line) and vertical (long-dashed line) components of the electric field across the channel of the transistor of Figure 3.6.

surface but is spread over a much wider region (Figure 3.8). The excess voltage represented by the difference between the applied drain voltage and $V_{DS,SAT}$ appears across this region (Figure 3.9), therefore leading to high electric fields. The vertical component of the electric field can be considerably higher than the linear regime discussed before. Additionally, the potential difference between gate and drain across a very short distance (i.e. the gate insulator) causes crowding of equipotential lines in the insulator near the drain junction. Figure 3.10 shows the electric field configuration; a distinct field gradient can be observed in that region. Because $V_{GS} < V_{DS}$ the electric field points towards the gate, therefore repelling electrons and attracting holes. Figure 3.11 shows the magnitude of those fields. If the field across the space charge region is high enough, avalanche multiplication can also occur.

This latter biasing condition has led to a lot of controversy in terms of explaining the injection mechanisms. Initially, it must be pointed out that the condition $V_{GS} < V_{DS}$ is ambiguous: depending on how much lower V_{GS} is, the effects upon device characteristics after degradation can be different.[12] This is because, under such conditions, the generation of both negative and positive charges take place, with the negative charges accumulating towards the centre of the channel and the positive ones near the drain junction. The most plausible explanation for this distribution is that hot electrons are injected even in the presence of a repelling field, provided that they have been re-directed towards the interface by phonon collisions.[5] As we have seen, the field is favourable to the injection of hot holes, but they are more difficult to inject due to the higher potential barrier they have to surpass. The gate bias can provide both the high vertical field which favours hot hole injection and can lead to avalanche conditions which can also contribute to the required hole heating. At the same time, this same vertical field repels electrons. Therefore there is a balance condition between the two effects; both hole and electron injection can take place and their intensity depends on the gate bias. The other source of disagreement amongst researchers comes from the different explanations given for the positive and negative channel charges present under such biasing conditions. In some cases the gate bias was not low enough to cause a predominance of positive charges and the situation was explained simply as hot electron injection and trapping.[13, 14] In other works the presence of both positive and negative charges has been attributed to hole

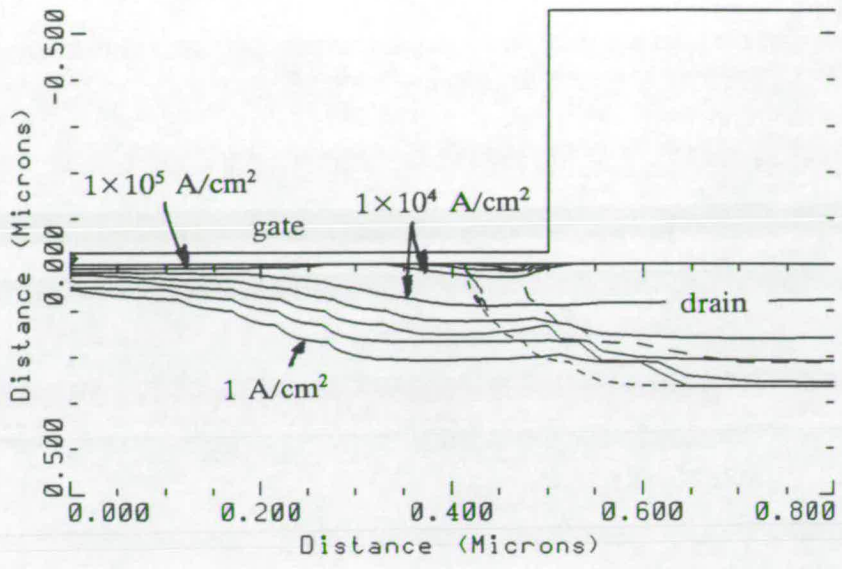


Figure 3.8: Simulated distribution of the electron current density (solid lines) near the channel of a 1μm n-channel transistor. $V_{GS}=1V$, $V_{DS}=5V$. The ratio between adjacent lines is 10.

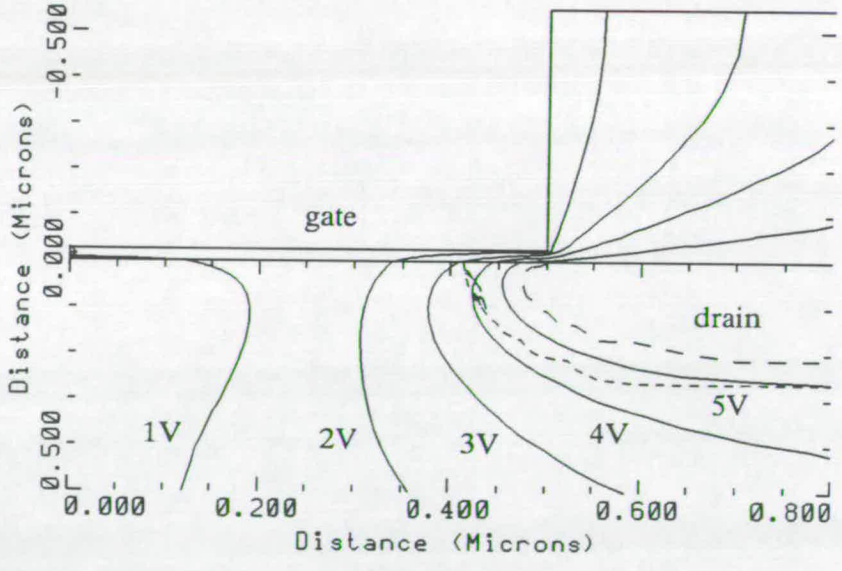


Figure 3.9: Potential distribution (solid lines) corresponding to the biasing condition for Figure 3.8. The increment between adjacent lines is 1V.

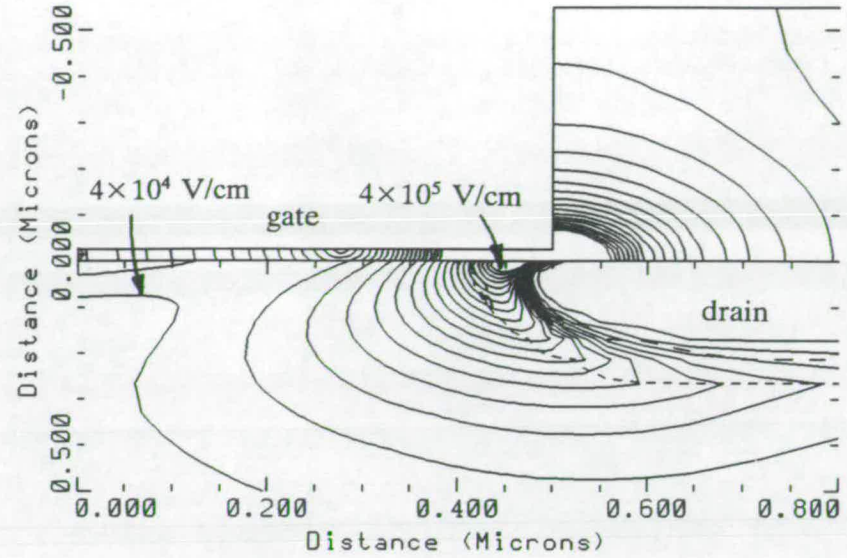


Figure 3.10: Electric field distribution (solid lines) corresponding to the biasing condition for Figure 3.8. The increment between adjacent lines is 2×10^4 V/cm.

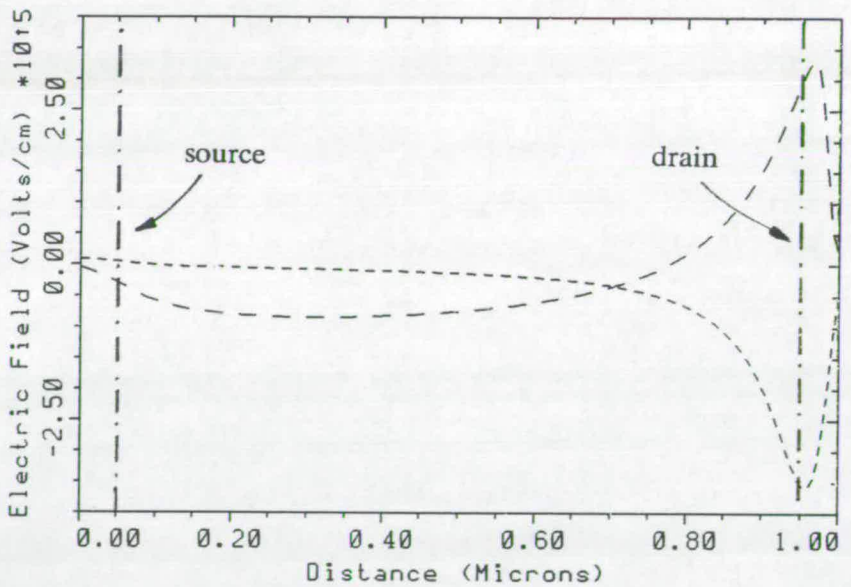


Figure 3.11: Horizontal (short-dashed line) and vertical (long-dashed line) components of the electric field across the channel of Figure 3.10.

injection, the negative charges being caused by the generation of acceptor-type interface states by the injected holes.[15, 16] The other source of discrepancy lies in the interpretation of the substrate current.[17] Holes generated by impact ionisation near the drain can either be heated and injected into the SiO_2 or simply flow to the bulk silicon, constituting the substrate current. The association of the maxima of parameter degradation and the substrate current has been seen as the maximum degradation being caused when the maximum number of hot holes has been generated.[18] The presence of such a current means that the electric field is high enough to cause impact ionisation, but the type of carrier being injected depends on local field distributions and on the impact ionisation process itself; therefore one cannot infer the type of injection taking place based solely on the nature of the substrate current.

3.4. Experimental methods

The investigation of hot carrier phenomena has been prompted by different objectives. Investigators who were devoted to the validation of generation and injection models,[10] and the study of trapping mechanisms,[19] amongst other objectives, needed a fairly high density of hot carriers and a uniform electric field, so they used either MOS capacitors or IGFETs operating in the substrate hot carrier generation mode. When the objective was the determination of the emission probability it became necessary to measure accurately both the generated and the injected currents, so substrate hot carrier generation was again used, this time with reentrant geometries (gate surrounding drain and source surrounding gate).[11, 20] The observation of degradation in IGFETs is difficult because it is a relatively slow process and the current constituted by the fraction of hot carriers which cause degradation cannot be directly measured. An indirect evaluation of this current can be obtained by measuring the carriers which are successful in crossing through the SiO_2 and constitute the gate current, but such measurement does not provide information on the ratio of those carriers which are effectively contributing towards aging. The substrate current has also been used to assess the density of heated carriers.[21] The maximisation of the substrate current has been extensively employed to accelerate hot carrier degradation tests.[17] However, as discussed in Section 3.3, the substrate current is nothing but a measure of the impact ionisation near the drain, which

is not a necessary condition for hot carrier generation.

Much research has been devoted to the investigation of the aging process, with three main objectives in mind: a) to find the real nature of the process, b) to determine device lifetime and c) to measure the magnitude and/or the energetic/spatial distribution of the interface state generation and charge trapping. Ultimately, the main objective has been aging minimisation and the improvement of structures.[17, 22] The methodology employed for the investigation of device aging has consisted of either looking at the device parameters and watching their variation under stressing conditions or the use of more sensitive techniques to anticipate parameter degradation (and try to avoid introducing additional aging by the test procedure itself). In this final section the more sensitive techniques which were used in the present work will be discussed, as well as the techniques used for determining the distribution (energetic and spatial) and magnitude of charge in the gate insulator.

3.4.1. The charge pumping method

The charge pumping effect was first observed by Brugler and Jespers [23] when they were evaluating enhancement MOSFETs operating under switching conditions and noticed the presence of a spurious substrate current. The set-up used is shown in Figure 3.12. Because the measured current was linearly proportional to the pulse frequency, they concluded that a fixed amount of charge was being pumped and measured at each pulse cycle. In order to further examine the effect, they applied voltage pulses to the gate with one level fixed at zero volts and the other varying. The resulting I_{CP} versus V_{GS} curve was observed to increase rapidly and then saturate, as depicted in Figure 3.13. The saturation voltage depended on the reverse bias applied to the junctions and corresponded to the values of threshold voltage (which also varies with the junction reverse bias due to the increase in surface potential [24]). It was also observed that for $V_{rev} = 0$ there is a small increase in I_{CP} even beyond the saturation point.

Their interpretation to the charge pumping effect was that when a voltage pulse is applied to the gate the channel is inverted, with the minority carriers being supplied by the adjacent reverse junctions. When the pulse is removed, the inversion layer vanishes and the minority carriers tend to go back to the junctions. However, some of them may

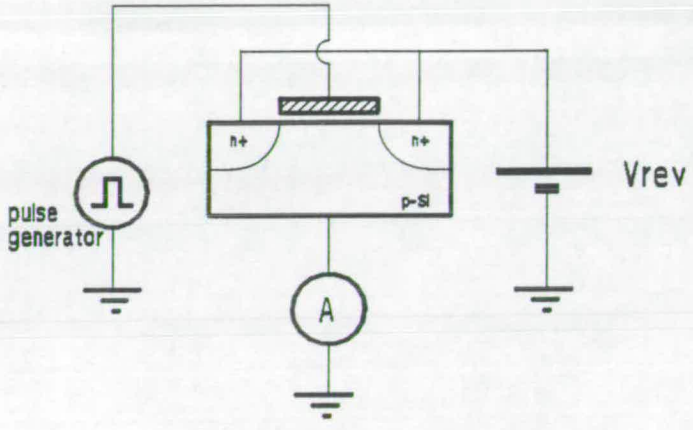


Figure 3.12: Measurement set-up for the charge pumping technique. Source and drain are reverse biased with the same voltage. A pulsed voltage is applied to the gate and the charge pumping current is measured as a substrate current.

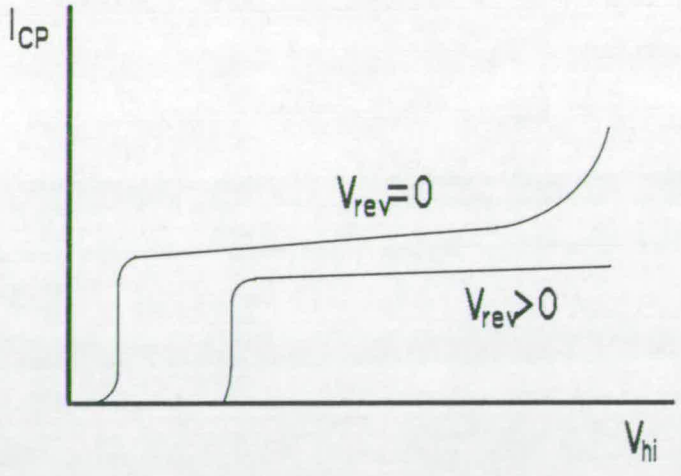


Figure 3.13: I_{CP} versus the upper level of the gate pulse with the lower level kept constant (Brugler's method).

recombine either in the bulk or at the surface. The majority carriers supplied for this recombination process by the substrate provide the charge pumping current. The total charge density involved in the recombination process can be expressed as

$$Q_T = \alpha Q_P + qN_{it} \tag{3.5}$$

where α is the fraction of carriers which suffer bulk recombination, Q_P is the charge density in the inversion layer ($\equiv -C(V_{GS} - V_T)$) and N_{it} is the density of interface states responsible for the surface recombination process. The resulting form for I_{CP} is

$$I_{CP} = A_G f [-\alpha C (V_{GS} - V_T) + qN_{it}] \tag{3.6}$$

where A_G is the effective channel area and f is the pulse frequency. The first term between square brackets is known as *geometric component* and the second *surface state component*. The geometric component is due to minority carriers which are situated far away from the junctions and therefore have to diffuse towards the substrate in order to find majority carriers to recombine. This is more common in long channel transistors when the rise and fall times of the gate pulse are too short (which implies that minority carriers have less time to drift back to the junctions). The geometric component is responsible for the small slope found beyond the saturation point in the I_{CP} versus V_{GB} curve because, unlike the surface state component, it is proportional to the charge density in the inversion layer.

The role of the interface states as the cause for the second and main component of the charge pumping current was confirmed by Brugler by inducing them under high field and high temperature conditions. Additionally, the density of interface states calculated from I_{CP} agreed with values normally expected for thermally grown SiO_2.

As the varying level of the gate voltage pulse is swept towards positive values (for an n-channel transistor) the bands bend downwards when observed from the bulk towards the interface. Interface states are then lowered below the Fermi level and are filled. With this voltage sweep the contribution of interface states situated at different energy levels are added to I_{CP}. However, this is limited by the threshold voltage, because beyond it the surface potential is "pinned" at approximately $2\Phi_F + V_{rev}$. Brugler and Jespers propose that this voltage sweep up to the saturation point of Figure 3.13 be used for the evaluation of the energy distribution of interface states. This technique is

disputed by Elliot,[25] who stated that when V_{GB} is varied below V_T the interface states whose energy levels are being scanned by the band bending cannot contribute for the charge pumping current. This happens because below threshold it is not possible for minority carriers to reach the channel region. Therefore he suggested a different technique for the determination of the energy distribution of interface states.

The new technique consisted of using pulses with constant amplitude and a varying d.c. offset. The resulting curve and the pulses corresponding to each one of its regions are depicted in Figure 3.14. If the sweep is set so that at one extreme both the upper and lower voltage level are in the inversion range and at the other extreme both levels are in accumulation, five distinct regions are observed:

a) when both levels are in accumulation (i.e. $V_{lo} < V_{hi} < V_{FB}$), no charge pumping current is generated since minority carriers are not being supplied to the channel because of the junction potential barrier;

b) when the high level is greater than the flatband voltage and lower than the threshold voltage ($V_{lo} < V_{FB} < V_{hi} < V_T$) minority carriers are increasingly being injected towards the channel during the high level and cause the charge pumping effect once the low level is reached;

c) if the high level is within strong inversion and the low level is at accumulation ($V_{lo} < V_{FB} < V_T < V_{hi}$) the full charge pumping current is observed;

d) as the low level starts depleting ($V_{FB} < V_{lo} < V_T < V_{hi}$) it becomes increasingly difficult for the minority carriers to drift back to the source/drain junctions and the charge pumping current starts to decrease;

e) when both the high and low levels are in the inversion range ($V_T < V_{lo} < V_{hi}$) it is no longer possible for the minority carriers to return to the junctions (the only charge pumping current still present is due to the geometric component).

The edge of the charge pumping curve corresponding to inversion (the leading one in the case of n-channel transistors) is dependent upon the reverse junction voltage, in the same way as the saturation point in the I_{CP} versus V_{GB} curve studied by Brugler and Jespers. On the other hand, Elliot stated that the other edge (corresponding to accumulation) can be used to compute the energy distribution of the interface states.

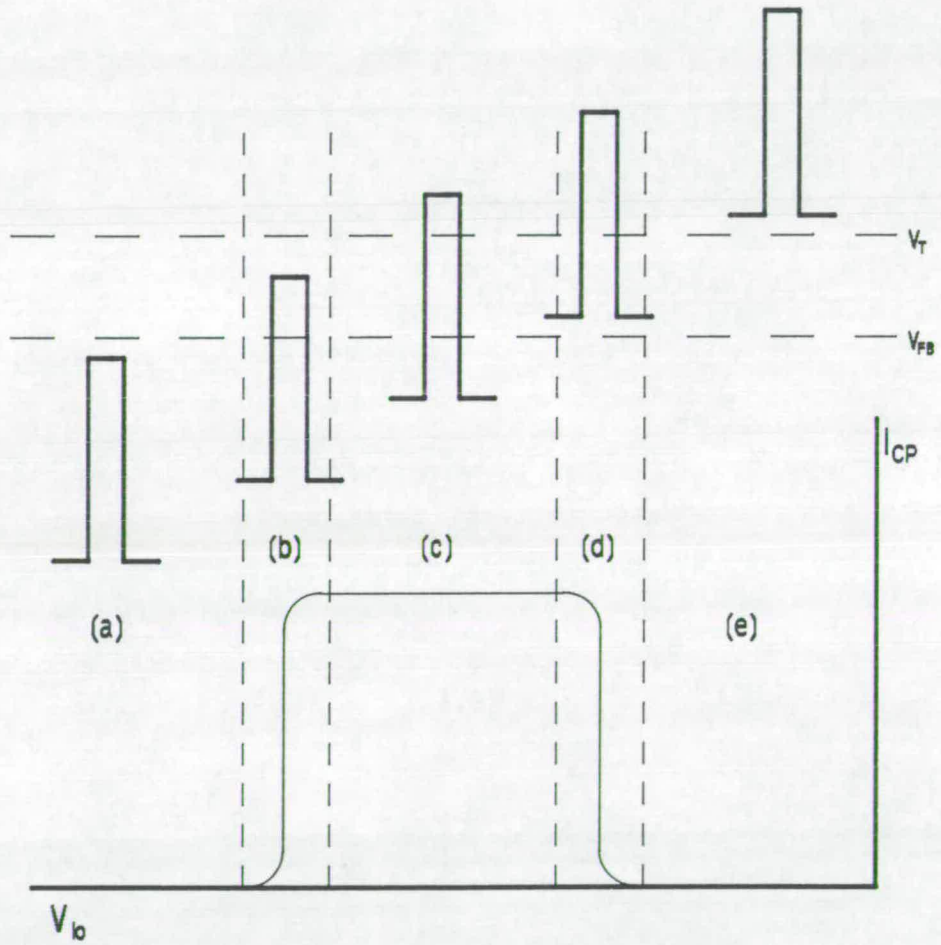


Figure 3.14: Elliot's technique for charge pumping measurement. Pulses with a constant amplitude are applied with an increasing d.c. offset. As the d.c. level is varied so that both levels are swept from accumulation to strong inversion, five distinct zones (marked "a" through "e") are observed.

A much better explanation for the charge pumping effect was given by Groeseneken,[26] who managed to explain the deviations observed in previous papers. For instance, he observed that the charge pumping current increases beyond the saturation point of the I_{CP} versus V_{GB} curve even when transition times of the gate pulse are increased or when the junctions are reverse biased. As the transition times are further increased, the charge pumping current decreases continuously instead of saturating. Therefore the hypothesis of the geometric component of the charge pumping current given by Brugler was no longer valid.

Groeseneken's model for the charge pumping mechanism was based on the carrier emission into the levels created by the interface states in the forbidden band. As the channel is driven into depletion by a gate voltage pulse, the bands bend downwards (considering an n-channel transistor) and the interface states are lowered into the Fermi level and are filled with valence electrons, initially coming from the valence band and, beyond threshold, coming from the conduction band. Depending on how fast the voltage is swept, the rate at which interface states become available can become greater than the rate of electron emission. This characterises a non-steady-state situation, in which the filling of the levels by electrons is only limited by the rate at which they are emitted. As threshold is approached, the surface concentration of electrons starts to increase significantly and the steady-state between the supply of electrons and the availability of interface states is restored. When the voltage is lowered again, electrons are emitted as more and more interface states are raised above the Fermi level. When the voltage becomes lower than V_T a non-steady-state situation between the rate at which electrons and states are available takes place again until, at flatband, the concentration of holes at the surface restores steady-state by making it easy for electrons to be injected from the states to the valence band. Electrons supplied or collected by the conduction band do so through the source/drain junctions; the emission or trapping of holes, on the other hand, give rise to a current of majority carriers in the bulk which can be measured as a substrate current; this current constitutes the charge pumping current. According to this model, the charge pumping effect involves the interface states situated between energies E_1 and E_2 within the bandgap.[27] These energies are given by

$$E_1 = E_i - kT \ln\{v_{th}\sigma_n n_i t_{em,e} + \exp[(E_i - E_{F,inv})/kT]\} \qquad (3.7.a)$$

$$E_2 = E_i + kT \ln\{v_{th}\,\sigma_p\, n_i\, t_{em,h} + \exp[(E_{F,acc} - E_i)/kT]\} \tag{3.7.b}$$

where v_{th} is the mean thermal velocity of electrons, σ is the capture cross section and $t_{em,e}$ and $t_{em,h}$ are the times during which non-steady-state emission of electrons and hole, respectively, take place. $t_{em,e}$ and $t_{em,h}$ are given by

$$t_{em,e} = t_{tra}\,\frac{|V_{FB} - V_T|}{|\Delta V_{GB}|} \tag{3.8.a}$$

$$t_{em,h} = t_{lea}\,\frac{|V_{FB} - V_T|}{|\Delta V_{GB}|} \tag{3.8.b}$$

where t_{tra} and t_{lea} are the times of the trailing and leading edges of the applied pulse. The final form for the charge pumping current then becomes

$$I_{CP} = 2qD_{it}fA_G kT\left[\ln(v_{th}\, n_i \sqrt{\sigma_n \sigma_p}) + \ln\left(\frac{|V_{FB} - V_T|}{|\Delta V_{GB}|}\sqrt{t_f\, t_r}\right)\right] \tag{3.9}$$

where D_{it} is the average density of interface states over the energies scanned by the Fermi level.

As we have seen before, Elliot proposed a dependence of the charge pumping current upon the surface potential. Groeseneken, on the other hand, presented a model based on the premise that the charge pumping current depends primarily on the times for the non-steady-state emission/trapping of carriers in the interface state. Observations such as the effect of the pulse transition times on the charge pumping current supported his model as opposed to Elliot's. Furthermore, according to Groeseneken and as we have discussed before, the emission/capture of valence electrons (hole trapping/emission) does not take place over the entire range from flatband to threshold; this makes Elliot's method invalid.

Groeseneken's method for the determination of the energy distribution of interface states is based on the measurement of the charge pumping current when the pulse transition times are varied. By varying independently the leading and trailing edges of the gate pulse it is possible to look at the energy distribution below and above midgap, respectively. The charge recombined in a pulse cycle can be written as

$$Q = qA_G \int_{E_1}^{E_2} D_{it}(E)\,dE \tag{3.10}$$

where E_1 and E_2 are the two energy boundaries defined by Equations (3.7.a) and (3.7.b). The final distribution was shown to be given by

$$D_{it}(E_1) = -\frac{t_{lea}}{qA_G kTf}\frac{dI_{CP}}{dt_{lea}} \tag{3.11.a}$$

$$D_{it}(E_2) = -\frac{t_{tra}}{qA_G kTf}\frac{dI_{CP}}{dt_{tra}} \tag{3.11.b}$$

The flatband and threshold voltages acquire a special meaning [27] in the charge pumping effect which differs from their classical definitions (as given, for example, by [24, 28, 29]). As we have seen in the explanation of Groeseneken's model above, the best way to describe the threshold voltage in view of the emission/trapping processes involved is the voltage at which the density of conduction electrons (in the case of an n-channel device) is large enough to fill all the interface states above the Fermi level during the time the pulse is applied. Similarly the flatband voltage is the voltage at which the density of holes in the valence band is large enough to fill the interface states during the time the pulse is removed. Despite this new definition for V_{FB} and V_T, the presence of charges in the oxide, which alter the surface potential, have the same effect of shifting those voltages as they would have in the case of their classical definitions. On the other hand, V_{FB} and V_T now strongly depend on the position across the channel, decreasing quite considerably near the source/drain junctions. This non-uniformity leads to one of the limitations of the charge pumping technique for the assessment of degradation phenomena: if injection is taking place too close to the junction it becomes difficult to be detected.

One of the advantages of the charge pumping technique in the analysis of non-uniform degradation is that the charge pumping curve provides a superimposed picture of the individual contributions of regions degraded differently.[27] In contrast with the measurement of device parameters (V_T, g_m, etc) which can only give a global view of the degradation process, this feature allows the observation of localised effects (Figure 3.15). In comparison with the charge pumping curve for the undegraded region, the degraded region will have its flatband and threshold voltages shifted by fixed charges or interface trapped charges to more positive (in the case of negative charges) or negative (in the case of positive charges) values and the contributions of both the undegraded and

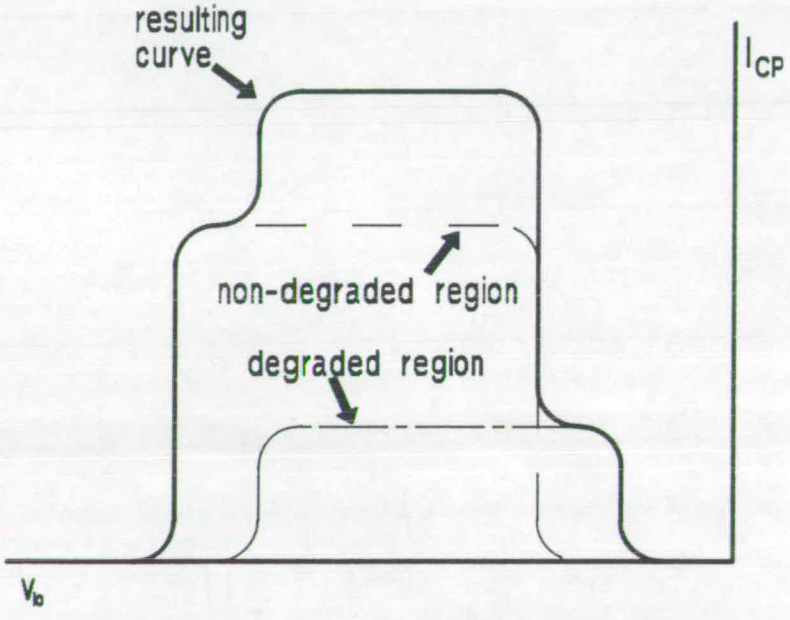


Figure 3.15: Charge pumping curve of a non-uniformly degraded transistor showing the contributions of the non-degraded and the degraded regions.

the degraded regions will be added to the total charge pumping current.

3.4.2. The gated diode method

Gated diode measurements were introduced by Grove in order to explain recombination/generation processes related to the surface space charge region.[30] When a reverse-biased gated diode has its gate biased at accumulation (Figure 3.16.a), only the generation/recombination centres within the junction depletion region will contribute to the reverse current (apart from any eventual leakage component). The generation current will be given by [29]

$$I_{acc} = \frac{qn_i}{2\tau_0} A_m x_d \tag{3.12}$$

where τ_0 is the lifetime, A_m is the area of the metallurgical junction and x_d is the width of the depletion region. As the gate voltage increases above flatband, the surface is depleted (Figure 3.16.b) and this new depletion region merges with the junction depletion region. Two new current components are then added to the reverse current: one introduced by this extended depletion region, which similarly to Equation 3.12 is given by

$$I_{depl} = \frac{qn_i}{2\tau_0} x_{ds} A_f \tag{3.13}$$

(where x_{ds} is the width of the depletion region under the gate and A_f is its area) and another component caused by the interface states and is given by

$$I_{int} = \frac{qn_i N_{it} v_{th} \sigma}{2} A_f . \tag{3.14}$$

The total reverse current therefore has a sharp increase as depicted in Figure 3.17). Because the first component (I_{depl}) depends on the width of the depletion region under the gate, which in turn is controlled by the gate voltage, the total reverse current continues to increase as the gate voltage is further increased. When the surface finally becomes inverted (Figure 3.16.c), generation at the interface states can no longer occur. Only the contributions of the depletion region surrounding the junction and under the gate remain, and the reverse current suffers a sharp drop (Figure 3.17).

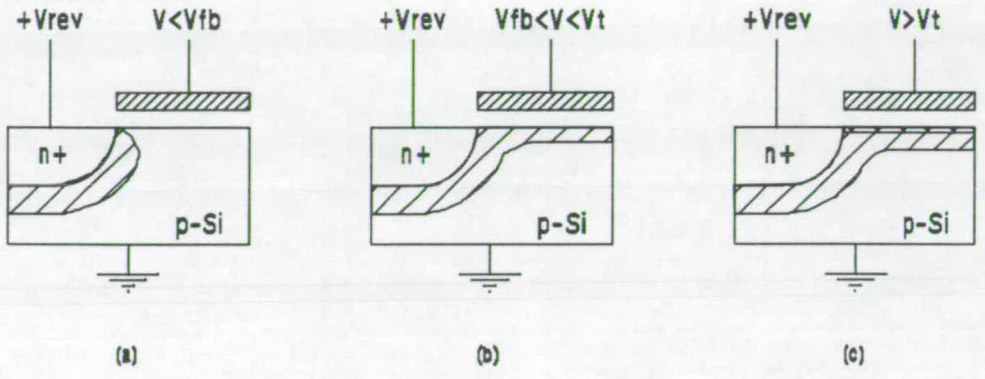


Figure 3.16: Cross section of a gated diode. (a) accumulation; (b) depletion; (c) inversion.

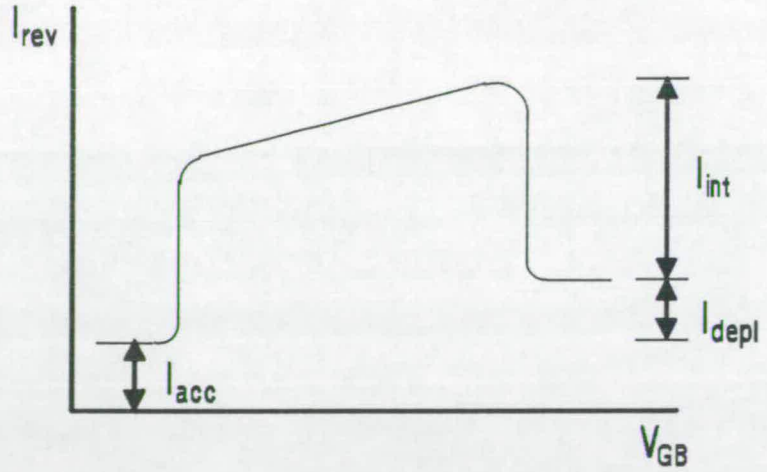


Figure 3.17: Gated diode curve (reverse current versus gate voltage) showing its components.

EX BIBL. UNIV. EDINBURGH

The use of the gated diode technique on an IGFET device in which one of the junctions is left floating has been proposed for degradation assessment by Giebel.[31] According to him, changes in the density of interface states can be detected by measuring the contribution of the I_{int} component (given by the difference between the maximum current before inversion and its value immediately after inversion). Any fixed or interface trapped charge would cause the flatband and threshold voltages to shift, thus affecting the slopes of the gated diode curve.

Giebel also states that the current corresponding to voltages immediately above flatband shows the contribution of the region where the junction depletion layer is overlapped by the gate This analysis is incorrect for the simple reason that at flatband the junction depletion region is already depleted by the junction reverse bias. Such misconceptions arise from the fact that, although there is no current flowing in the channel of a gated diode, the region is not under thermal equilibrium. A gated diode with the surface under the flatband condition can be seen as an ordinary non-gated diode. When a reverse bias is applied, the junction is no more under thermal equilibrium and the Fermi level is split into the electron and hole quasi-Fermi levels (also known as *imrefs*). Beyond the depletion layer the semiconductor is still under thermal equilibrium and the imrefs are joined again. When the surface is driven into depletion by the gate bias, the density of minority carriers at the surface tends to increase. However, the reverse-biased junction will drain those carriers. This decrease in the concentration of minority carriers will not be compensated by any other source, which means that the imrefs will not be joined again throughout the whole length of the channel. This non-thermal-equilibrium situation is totally different from the standard MOS capacitor, in which the concentration of minority carriers in the surface depletion layer depends uniquely on the surface potential and the equilibrium minority carrier density.

In comparison with the charge pumping technique, the gated diode method is restricted to the observation of interface states near the midgap region, with no possibility of scanning the energy levels. This will be further discussed in Chapter 6. Nevertheless, it provides an accurate way to measure V_{FB} and V_T in their classical definitions, therefore not being subject to local variations of carrier density as in the case of the charge pumping method.

References

1. Elliott, S. R., *Physics of Amorphous Materials,* Longman, Essex, 1990.
2. Deal, B. E., "Standardized terminology for oxide charges associated with thermally oxidized silicon," *IEEE Trans. Electron Devices*, vol. ED-27, pp. 606-608, March 1980.
3. Ghandhi, S. K., *VLSI Fabrication Principles,* Wiley, New York, 1983.
4. Warner, R. M. and Grung, B. L., *Transistors,* Wiley, New York, 1983.
5. Barbottin, G. and Vapaille, A., *Instabilities in Silicon Devices,* Elsevier, (North Holland), 1986.
6. Tam, S., Ko, P. K., and Hu, C., "Lucky-electron model of channel hot-electron injection in MOSFETs," *IEEE Trans. Electron Devices*, vol. ED-31, pp. 1116-1125, September 1984.
7. Ning, T. H., "Hot-electron emission from silicon into silicon dioxide," *Solid-State Electron.*, vol. 21, pp. 273-282, 1978.
8. Takeda, E., Kume, H., Toyabe, T., and Asai, S., "Submicrometer MOSFET structure for minimizing hot-carrier generation," *IEEE Trans. Electron Devices*, vol. ED-29, pp. 611-618, April 1982.
9. Cottrell, P. E., Troutman, R. R., and Ning, T. H., "Hot-electron emission in n-channel IGFETs," *IEEE Trans. Electron Devices*, vol. ED-26, pp. 520-533, April 1979.
10. Verwey, J. F., Kramer, R. P., and de Maag, B. J., "Mean free path of hot electrons at the surface of boron-doped silicon," *J. Appl. Phys.*, vol. 46, pp. 2612-2619, June 1975.
11. Ning, T. H., Osburn, C. M., and Yu, H. N., "Emission probability of hot electrons from silicon into silicon dioxide," *J. Appl. Phys.*, vol. 48, pp. 286-293, January 1977.
12. Hofmann, K. R., Werner, C., Weber, W., and Dorda, G., "Hot-electron and hole-emission effects in short n-channel MOSFETs," *IEEE Trans. Electron Devices*, vol. ED-32, pp. 691-699, March 1985.

13. Chen, K. L., Saller, S. A., Groves, I. A., and Scott, D. B., "Reliability effects on MOS transistors due to hot-carrier injection," *IEEE J. Solid-State Circuits*, vol. SC-20, pp. 306-313, February 1985.

14. Ng, K. K. and Taylor, G. W., "Effects of hot-carrier trapping in n- and p-channel MOSFETs," *IEEE Trans. Electron Devices*, vol. ED-30, pp. 871-876, August 1983.

15. Fair, R. B. and Sun, R. C., "Threshold-voltage instability in MOSFETs due to channel hot-hole emission," *IEEE Trans. Electron Devices*, vol. ED-28, pp. 83-94, January 1981.

16. Gesch, H., Leburton, J. P., and Dorda, G. F., "Generation of interface states by hot hole injection in MOSFETs," *IEEE Trans. Electron Devices*, vol. ED-29, pp. 913-918, May 1982.

17. Hu, C., Tam, S. C., Hsu, F. C., Ko, P. K., Chan, T. Y., and Terrill, K. W., "Hot electron-induced MOSFET degradation - model, monitor, and improvement," *IEEE J. Solid-State Circuits*, vol. SC-20, pp. 295-305, February 1985.

18. Takeda, E., Shimizu, A., and Hagiwara, T., "Role of hot-hole injection in hot-carrier effects and the small degraded channel region in MOSFETs," *IEEE Electron Device Lett.*, vol. EDL-4, pp. 329-331, September 1983.

19. Aitken, J. M. and Young, D. R., "Avalanche injection of holes into silicon dioxide," *IEEE Trans. Nucl. Sci.*, vol. NS-24, pp. 2128-2134, December 1977.

20. Troutman, R. R., "Silicon surface emission of hot electrons," *Solid-State Electron.*, vol. 21, pp. 283-289, 1978.

21. Tam, S., Hsu, F. C., Hu, C., and Muller, R. S., "Correlation between substrate and gate currents in MOSFETs," *IEEE Trans. Electron Devices*, vol. ED-29, pp. 1740-1744, November 1982.

22. Mikoshiba, H., Horiuchi, T., and Hamano, K., "Comparison of drain structures in n-channel MOSFETs," *IEEE Trans. Electron Devices*, vol. ED-33, pp. 140-144, January 1986.

23. Brugler, J. S. and Jespers, P. G. A., "Charge pumping in MOS devices," *IEEE Trans. Electron Devices*, vol. ED-16, pp. 297-302, March 1969.

24. Grove, A. S., *Physics and Technology of Semiconductor Devices,* Wiley, New York, 1967.

25. Elliot, A. B. M., "The use of charge pumping currents to measure surface state densities in MOS transistors," *Solid-State Electron.*, vol. 19, pp. 241-247, 1976.

26. Groeseneken, G., Maes, H. E., Beltran, N., and De Keersmaecker, R. F., "A reliable approach to charge-pumping measurements in MOS transistors," *IEEE Trans. Electron Devices*, vol. ED-31, pp. 42-53, January 1984.

27. Heremans, P., Witters, J., Groeseneken, G., and Maes, H. E., "Analysis of the charge pumping technique and its application for the evaluation of MOSFET degradation," *IEEE Trans. Electron Devices*, vol. 36, pp. 1318-1335, July 1989.

28. Sze, S. M., *Physics of Semiconductor Devices,* Wiley, New York, 1981.

29. Muller, R. S. and Kamins, T. I., *Device Electronics for Integrated Circuits,* Wiley, New York, 1986.

30. Grove, A. S. and Fitzgerald, D. J., "Surface effects on p-n junctions: charateristics of surface space-charge regions under non-equilibrium conditions," *Solid-State Electron.*, vol. 9, pp. 783-806, 1966.

31. Giebel, T. and Goser, K., "Hot-carrier degradation of n-channel MOSFETs characterized by a gated-diode measurement technique," *IEEE Electron Device Lett.*, vol. 10, pp. 76-78, February 1989.

Chapter 4 - Device Fabrication

4.1. Introduction

The unusual characteristics of the transistors used in the present work, more particularly the presence of transistors with offset gates with respect to the transistor channel, required special fabrication procedures. Nevertheless, the bulk of the fabrication process remained essentially the same as any conventional silicon gate nMOS process. The discussion of standard processes would be rather inappropriate here, for it would be lengthy and unnecessary in view of the number of publications available on the subject. However, the variations which had to be introduced in such processes are most certainly worth discussing; they led, in some cases, to very interesting results and to a better understanding of the processes themselves and the mechanisms involved. Unfortunately the problems encountered, as well as many interesting phenomena which were observed, could only be tackled at a very superficial level in view of the fact that fabrication was not the main issue to be addressed. Much of it is, therefore, left as a suggestion for future work.

Two different processes were used in order to produce the required samples. Figure 4.1 summarises the strategy for experiments and and the decision-making involved in the processes. The first process led to arrays of self-aligned transistors with a range of channel lengths and widths and different arrays of 1.0 and 1.5 μm offset-gate transistors with and without lightly doped drain (LDD). The self-aligned transistors presented severe junction leakage and breakdown problems. The offset-gate LDD transistors, whose LDD regions were defined by photolithography, presented difficulties in the assessment of LDD dimensions as it will be further discussed in the next Chapter. This problem, together with the quest for smaller transistors, led to a second process. This new process employed a totally different mask design, which included offset-gate transistors only. In this second batch, the LDD regions were created by angled implants using

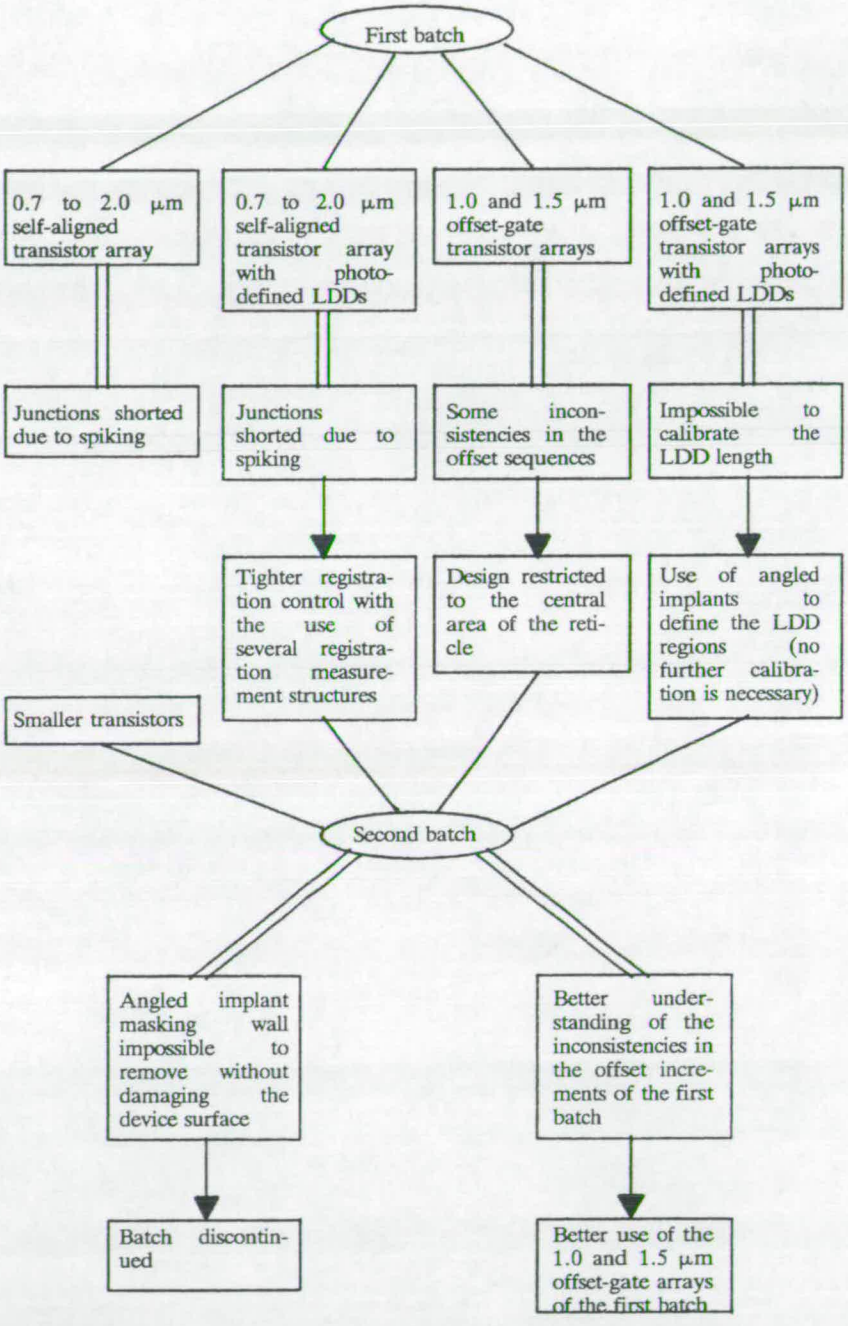


Figure 4.1: Strategy used for the design and fabrication in this project (a '—' line means 'features', a '=' line means 'results' and a '→' means 'consequence')

implant masking walls. These walls were extremely difficult to remove without severely affecting device performance, so the second batch had to be discontinued. Nonetheless, it allowed a much better understanding of the alignment problems that occurred in the previous batch. Additionally, very interesting phenomena were observed which had not been noticed before, particularly insofar as multilayer dry etching and the wet etching of heavily doped silicon is concerned.

4.2. First fabrication

4.2.1. Design considerations

The objective of the first design used in this project was to produce arrays of transistors in which different geometric variables could be varied independently. Five different arrays were chosen:

1) Self-aligned transistors with variable channel length, ranging from 2.0 μm to the minimum achievable channel length, which in this case was 0.8 μm.

2) Repetitions of the array above with a shallow implant overlapping the channel in one of the source/drain sides, thus creating an asymmetrical LDD structure. This array is repeated with 0.1, 0.2 and 0.3 μm overlaps.

3) Self-aligned transistors with variable channel width, ranging from 5.0 μm to 1.0 μm.

4) Non-self-aligned transistors, in which the gate can be placed independently from the channel position. A systematic variation in gate offset and in gate length takes place across each row and column of the array, respectively. Gate offsets vary from +0.3 to -0.3 μm with respect to the channel position and gate lengths vary from 1.0 to 1.5 μm (both variations in 0.1 μm steps), therefore resulting in a range of gaps and overlaps varying in 0.05 μm steps. The array is divided into two sub-arrays, one in which all transistors have 1.0 μm channels and another with 1.5 μm channels. Figure 4.2 shows a simplified diagram of such array.

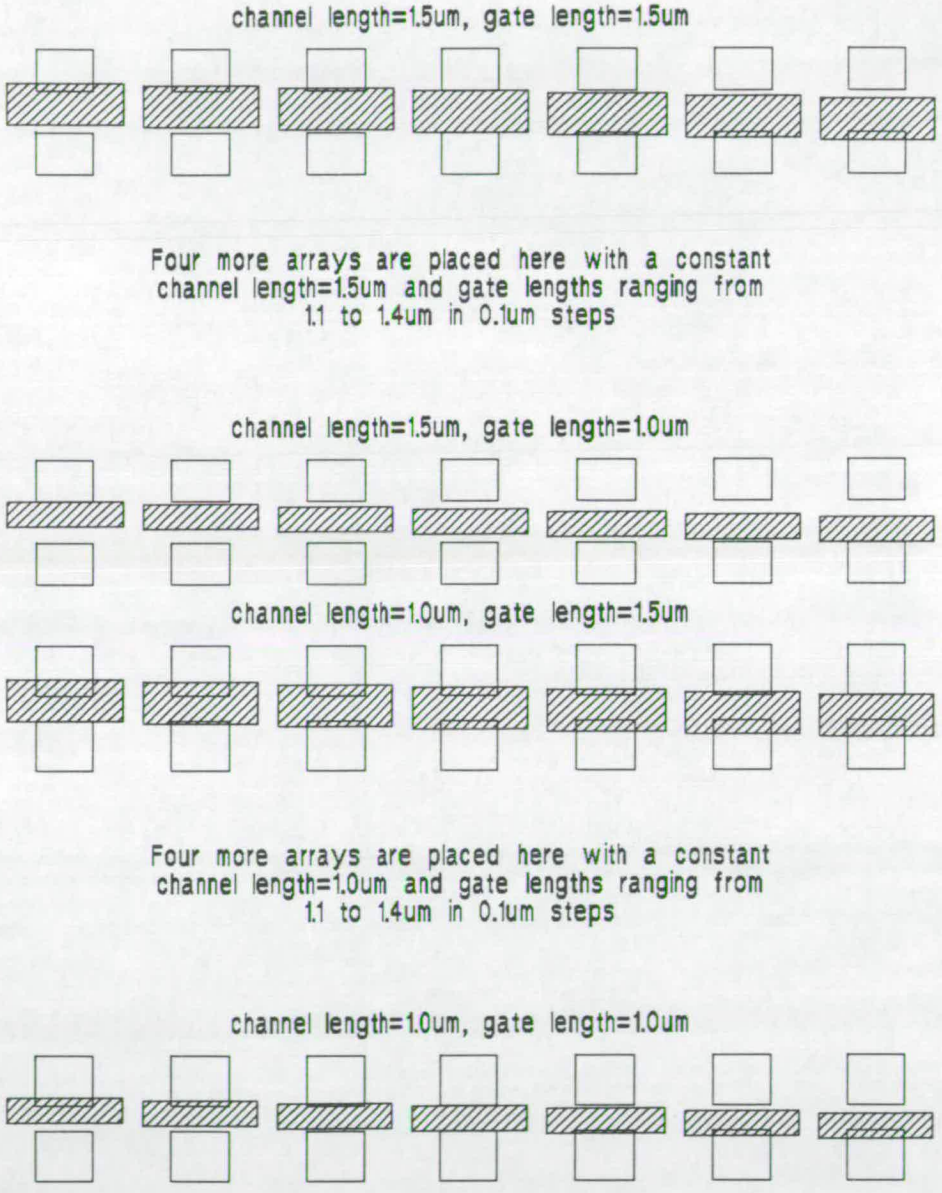


Figure 4.2: Array of offset-gate transistors used in the first design. The same array is repeated with LDD transistors, each array with a different LDD overlap (0.1, 0.2, 0.3 μm overlap on the drain side only and 0.2 and 0.3 μm on both source and drain sides).

5) Repetitions of the array above with a shallow implant overlapping the channel in one of the source/drain sides, thus creating an asymmetrical LDD structure. The array is repeated with 0.1, 0.2 and 0.3 μm overlaps.

6) Repetitions of the array 5, this time with shallow implant overlaps on both sides. Two different overlap sizes were included: 0.2 and 0.3 μm.

The 0.1 μm increments are produced by 1 μm offsets in the reticle since a 10:1 stepper was used. Within each exposure field a random offset of up to ±0.3 μm may be introduced by the alignment process and this offset is superimposed on all transistors in a chip. A single calibration of this registration offset can therefore be applied to all devices simultaneously to give an accurate value of the gap size for each device. However, as will be seen in Section 4.2.2.1, distortions also occur within each field and they must be calibrated at every transistor site.

The layout was subdivided into four main chips: chip one with all the self-aligned arrays described in items 1, 2 and 3 above; chip two with the non-LDD, offset gate transistor array and the 0.1 μm LDD overlap array; chip three with the 0.2 and 0.3 μm LDD overlap arrays; and chip 4 with the arrays described in item 6 above. This chip division was done in order to make it easier to encapsulate devices in a later stage of the work.

In order to avoid that too large deviations in gate position and LDD overlap were introduced by the photolithographic process, a monitoring technique had to be adopted. Vernier structures such as the depicted in Figure 4.3 were included in three of the four chips. Additionally, special transistors (Figure 4.4) were included in chip one with the specific purpose of assessing registration, in an attempt to reproduce electrically the results read with the aid of the vernier structures. Standard alignment marks were also included both on the chips (Figure 4.5) and at the sides of each wafer (Figure 4.6) for the purpose of global alignment.

The mask set used the following layers:

Layer 1: active areas

Layer 2: shallow implant (for the definition of the LDD areas)

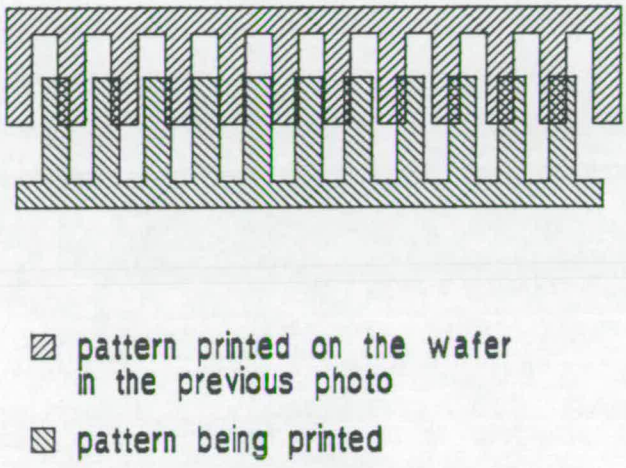


Figure 4.3: Vernier structure for assessing registration.

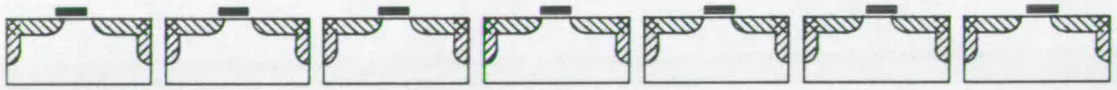


Figure 4.4: Alignment transistors. They have extended LDD regions to allow the verification of the registration between the gate and the LDD levels. This, allied to the measurements obtained from non-LDD offset-gate transistors, was expected to provide information on the relative misalignments among source/drain, LDD and gate levels.

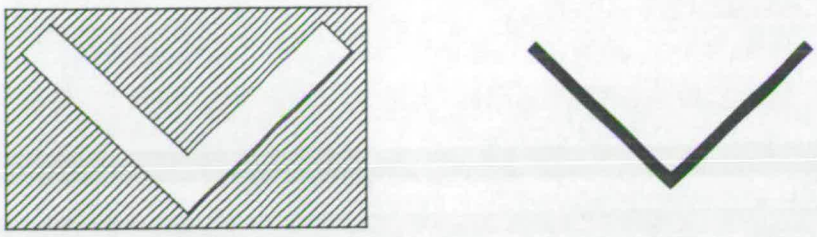


Figure 4.5: On-site alignment mark. The mark on the right-hand side is printed on the wafer. In a subsequent photolithographic step it is centred with respect to the diagonal windows of the pattern in the left-hand side which is present in the next mask level.

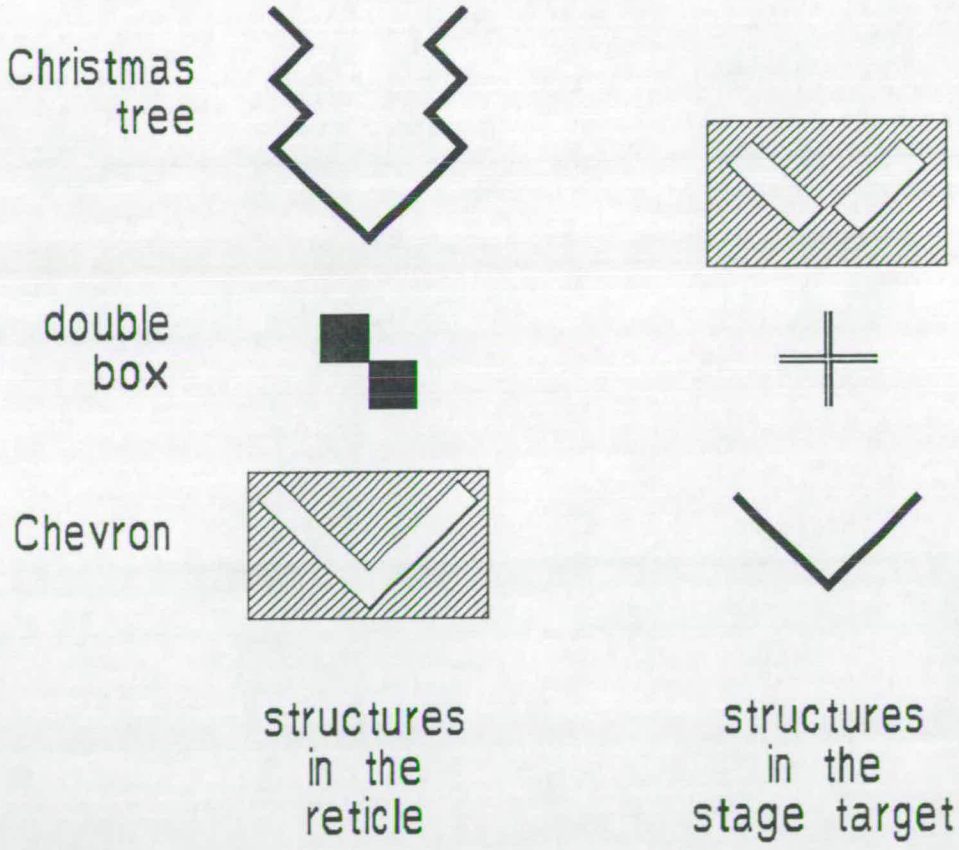


Figure 4.6: Global alignment marks. These marks are used for coarse alignment between reticle and stage and between wafer and the stepper's fixed targets.

Layer 3: source/drain definition of the offset gate transistors

Layer 4: first gate cut (for the self-aligned transistors)

Layer 5: second gate cut (for the offset gate transistors)

Layer 6: contact windows

Layer 7: aluminium tracks

Layer 8: passivation windows

The alignment sequence adopted is depicted in Figure 4.7.

4.2.2. Fabrication process

The process employed was based on the EMF standard 1.5 μm nMOS polysilicon gate process. Some modifications were necessary in order to produce offset gate transistors whilst also including standard self-aligned transistors. Figure 4.8 illustrates the integration of both types of transistors in the same process. The description of the whole process is depicted in Figure 4.9. Additionally, a tighter control over photolithographic tolerances was also included. Due to the presence of very shallow junctions, the contact implant also had to be very carefully analysed.

4.2.2.1. Alignment control

The assessment technique chosen for measuring the gate offsets, which will be discussed in detail in Chapter 5, requires a sequence of increasing offsets in which a zero offset transistor is also included. Considering that the drawn offsets range from 0 to 0.3 μm on both directions (i.e., -0.3 to +0.3 μm), the registration would have to be kept within those limits. However, as can be noticed in Figure 4.7, there was a risk that errors at different photolithographic steps could accumulate, for in some cases a given layer would not be matched to that immediately preceding. For example, layer 4 matches to layer 1, so the registration errors from layers 3 and 4 would eventually add up in terms of gate positioning with respect to the channel. This is depicted in Figure 4.10. Therefore all photolithographic steps were carefully monitored during fabrication with the aid of the vernier structures.

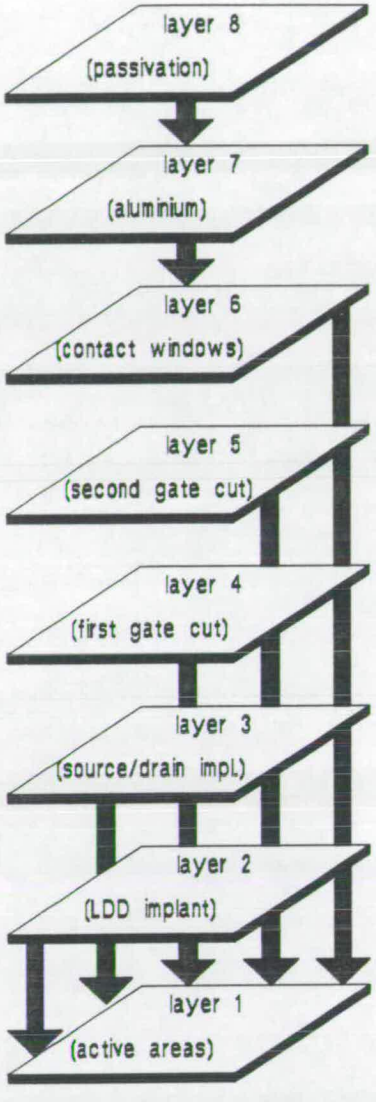


Figure 4.7: Alignment sequence for the first fabrication.

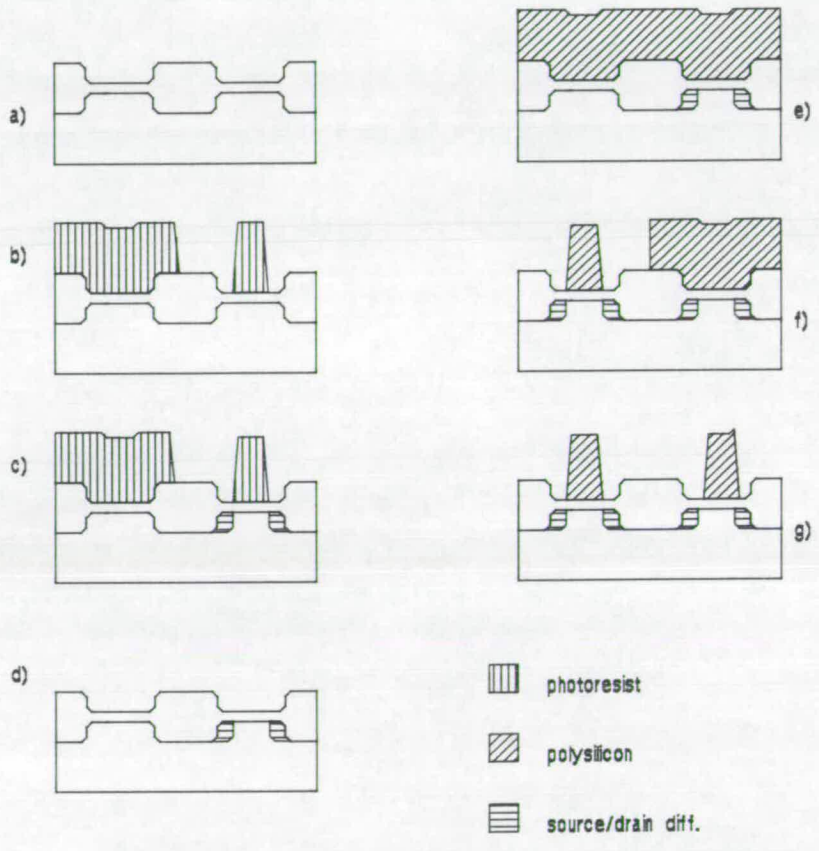
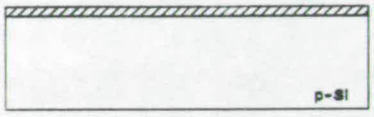
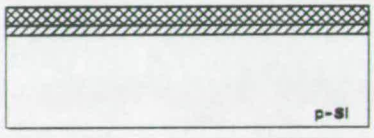


Figure 4.8: Process integration of self-aligned and offset-gate transistors.

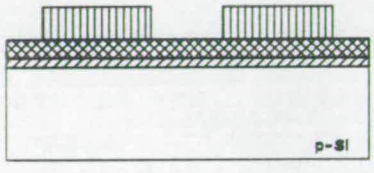
Initial oxidation (5 minutes O_2+HCl / 5 minutes H_2+O_2+HCl / 5 minutes O_2 at 950°C)
Target thickness = 400 Å



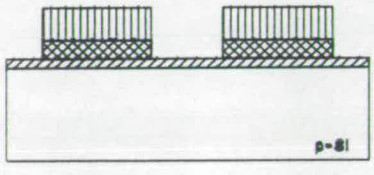
Silicon nitride deposition (20 minutes in SiH_2Cl_2+NH_3 at 200 mTorr, 850°C)
Target thickness = 1000 Å



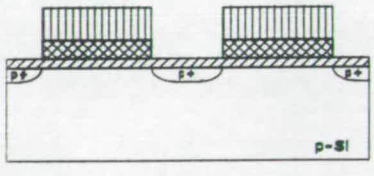
First photo (active areas)



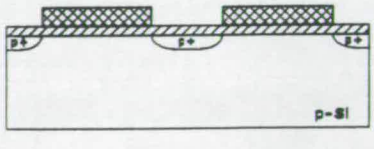
Si_3N_4 reactive ion etch - 30 minutes in CF_4:H_2 plasma



Boron implant (isolation implant) (dose=7×10^{12} ions/cm^2, energy = 70 keV)



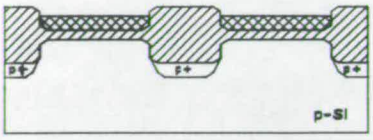
Photoresist removal with fuming HNO_3



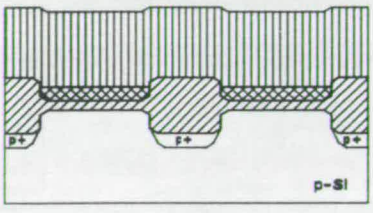
Pre-diffusion clean - 15 seconds in 10% HF

Figure 4.9: Description of the first fabrication.

Field oxidation (5 minutes O_2+HCl / 30 minutes H_2+O_2+HCl / 3 hours H_2+O_2 / 5 minutes O_2 at 950°C)
Target thickness = 6000 Å

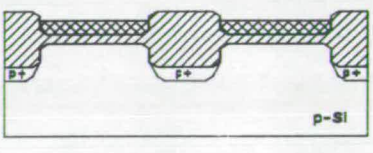


Photoresist coat (to protect the front of the wafers)



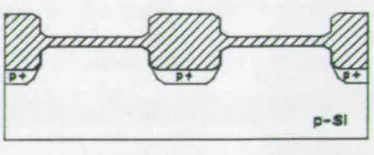
Back etch - Immerse in 4:1 NH_4F:HF until the backs dewet

Photoresist removal with fuming HNO_3



Oxide etch (to remove the top oxynitride) - 15 seconds in 4:1 NH_4F:HF

Si_3N_4 wet etch - 30 minutes in H_3PO_4 at 165°C



Oxide etch (to remove the pad oxide from the active areas) - Immerse in 4:1 NH_4F:HF until the scribe lines dewet

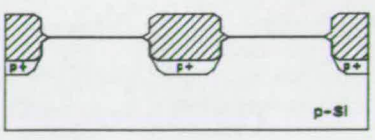
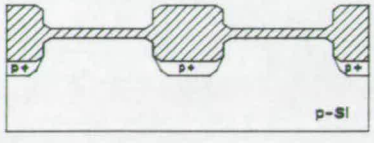


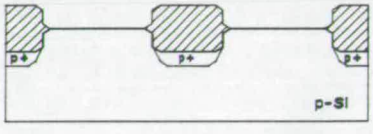
Figure 4.9: Description of the first fabrication (cont'd).

RCA clean - 10 minutes in 1:1:5 NH_3:H_2O_2:H_2O at 80°C / wash in DI water / 10 minutes in 1:1:5 HCl:H_2O_2:H_2O at 80°C / wash in DI water

Sacrificial oxidation (5 minutes O_2+HCl / 5 minutes H_2+O_2+HCl/5 minutes O_2 at 950°C



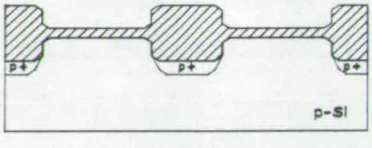
Oxide etch (to remove sacrificial oxide) - Immerse in 4:1 NH_4F:HF until the backs dewet



Karo's clean - 10 minutes in boiling 2:1 H_2SO_4:H_2O_2 (5 wafers only)

RCA clean - 10 minutes in 1:1:5 NH_3:H_2O_2:H_2O at 80°C / wash in DI water / 10 minutes in 1:1:5 HCl:H_2O_2:H_2O at 80°C / wash in DI water

Gate oxide (5 minutes O_2+HCl / 2 minutes H_2+O_2+HCl / 5 minutes O_2 at 950°C)
Target thickness = 250 Å

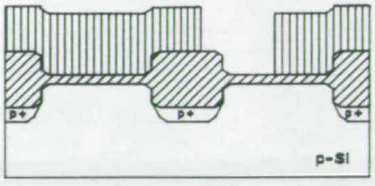


Boron implant (threshold adjust and punchthrough control)
(dose=7×10^{11} ions/cm^2, energy = 25 keV
dose=2×10^{12} ions/cm^2, energy = 140 keV)

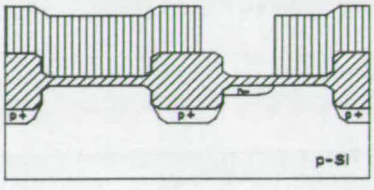
RCA clean - 10 minutes in 1:1:5 NH_3:H_2O_2:H_2O at 80°C / wash in DI water / 10 minutes in 1:1:5 HCl:H_2O_2:H_2O at 80°C / wash in DI water

Figure 4.9: Description of the first fabrication (cont'd).

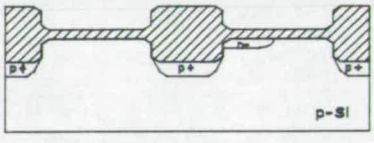
Second photo (LDD definition) - uses high energy photoresist



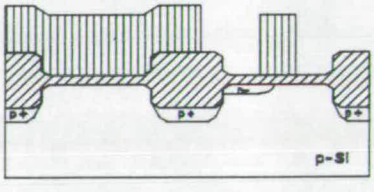
Arsenic implant (LDD implant) (dose=4×10^{13} ions/cm^2, energy = 50 keV)



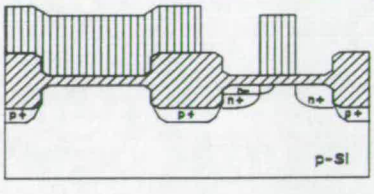
Photoresist removal



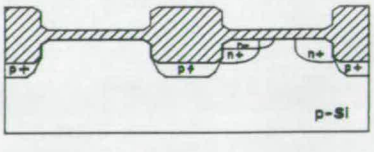
Third photo (source/drain definition) - uses high energy photoresist



Arsenic implant (source/drain implant of the offset transistors) - dose=7×10^{15} ions/cm^2, energy = 90 keV



Photoresist removal



Polysilicon deposition (40 minutes in SiH_4 at 200 mTorr, 600°C)
Target thickness = 4000 Å

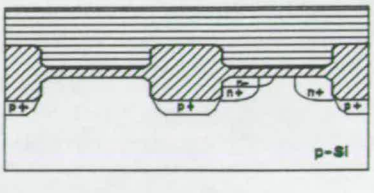
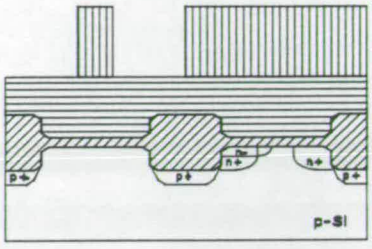
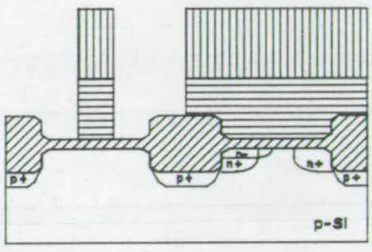


Figure 4.9: Description of the first fabrication (cont'd).

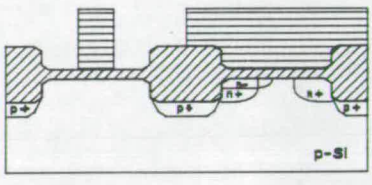
Fourth photo (first gate patterning)



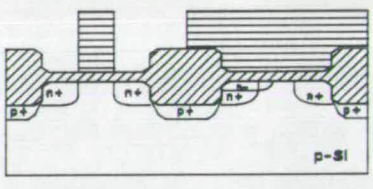
Polysilicon reactive ion etch - 10 minutes in CCl_4:N_2 plasma



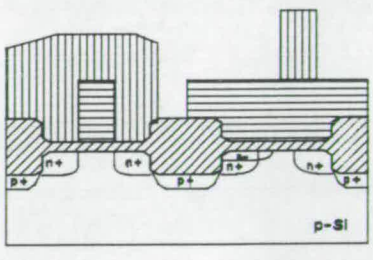
Photoresist removal



Arsenic implant (source/drain implant) dose=7×10^{15}ions/cm^2, energy = 90 keV



Fifth photo (second gate patterning)



Polysilicon reactive ion etch - 10 minutes in CCl_4:N_2 plasma

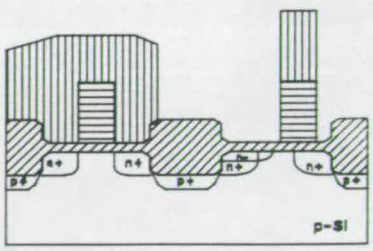
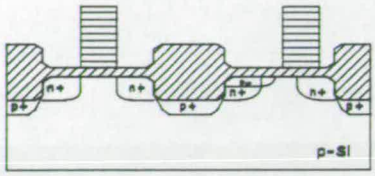
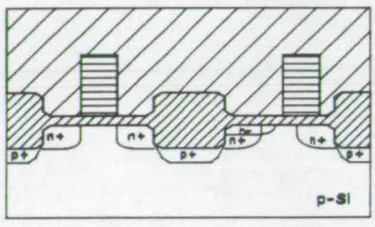


Figure 4.9: Description of the first fabrication (cont'd).

Photoresist removal



Pyrolitic deposition of SiO_2 (initial passivation)
SiH_4:O_2 reaction on a hotplate set to 430°C

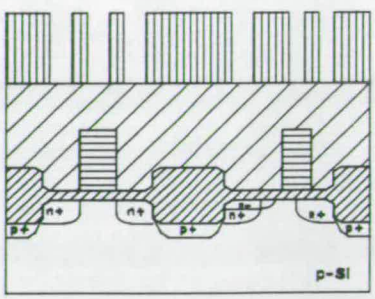


Pyrolitic deposition of phosphorus-doped SiO_2
SiH_4:O_2 reaction on a hotplate set to 430°C
(with phosphine added)

Reflow of the pyrolytic SiO_2 - 10 minutes in an oven set to 1000°C with nitrogen flowing

Densification of the pyrolitic SiO_2 (5 minutes O_2 /
2 minutes H_2+O_2 / 5 minutes O_2 at 950°C)

Sixth photo (contact windows)



SiO_2 reactive ion etch - 40 minutes in CHF_3:He plasma

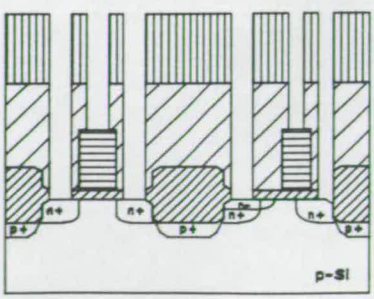
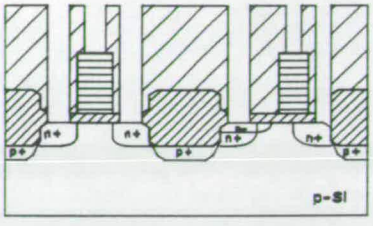


Figure 4.9: Description of the first fabrication (cont'd).

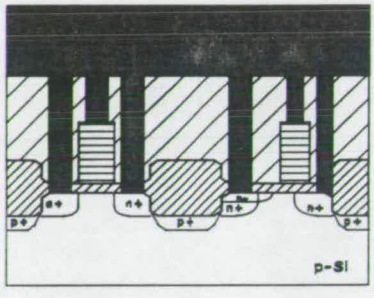
Photoresist removal



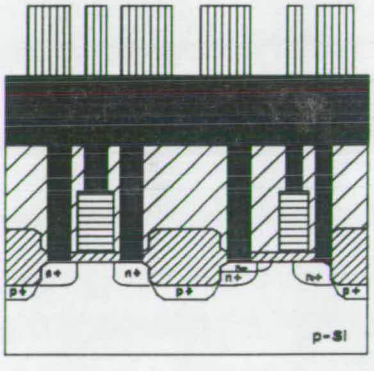
Phosphorus implant (spiking prevention)
dose= 1×10^{13} ions/cm^2, accelerating potential = 90 kV
with doubly ionised species resulting in an energy of 180 keV

Pre-metallisation clean - 20 seconds in 25:1 NH_4F:HF

Aluminium sputtering - use of aluminium with 2% silicon
Target thickness = 1 μm



Seventh photo (metal patterning)



Metal etch - 5 minutes in CCl_4:N_2 plasma

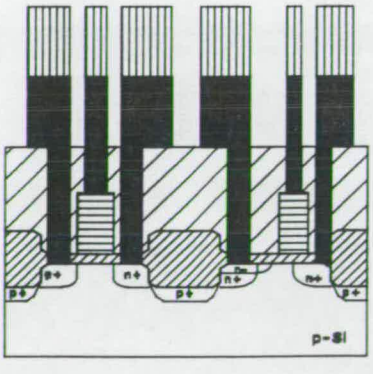
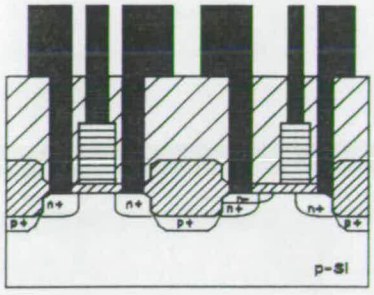


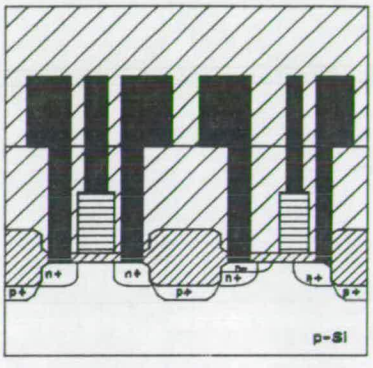
Figure 4.9: Description of the first fabrication (cont'd).

Photoresist removal

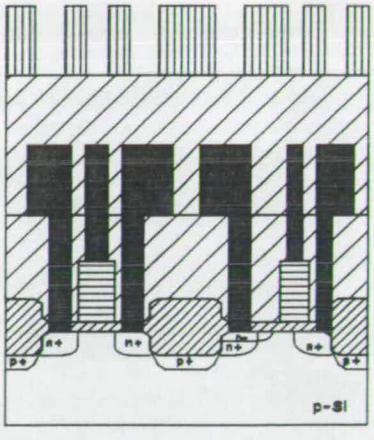


Contact sintering (5 minutes N_2 / 10 minutes H_2+N_2 / 5 minutes N_2 at 435°C)

Pyrolitic deposition of SiO_2 (final passivation) SiH_4:O_2 reaction on a hotplate set to 430°C



Eighth photo (pad window cut)



Pad window etch - Immerse in 4:1 NH_4F:HF until scribe lines dewet

Photoresist removal

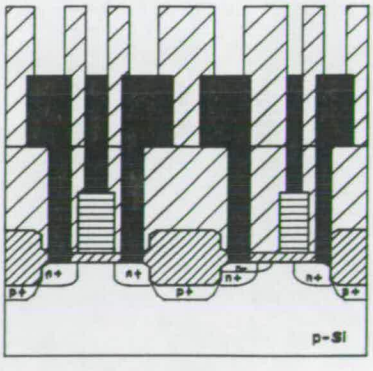


Figure 4.9: Description of the first fabrication (cont'd).

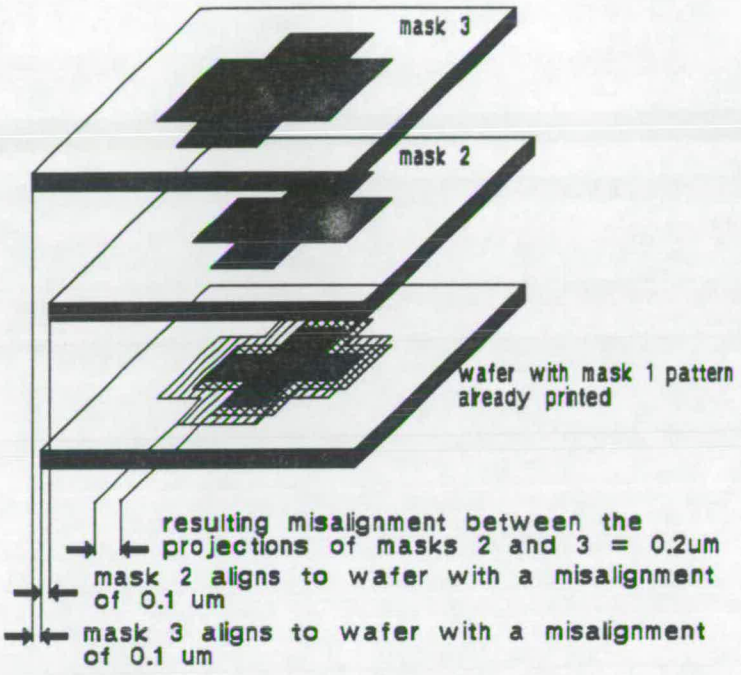


Figure 4.10: Example of accumulation of alignment errors.

At first, verniers were "manually" read using an ordinary microscope. However, this proved to be an extremely demanding task considering that there were three readings to be done in each group of four chips, 38 of such groups in a wafer and ten wafers in the initial batch, bringing the measurements to a total of 1,140 measurements for every step or trial run performed. Such a number of repetitive measurements puts at risk the very reliability of the results and causes unnecessary delays in the overall processing time. An automated technique was then introduced which made use of a Quaestor programmable microscope. This instrument also had the advantage of recording all the results for later use. The automated measurement of a vernier is depicted in Figure 4.11. Notice that only the central portion of the vernier is actually used (unlike "manual" measurements, there is no need in this case to look at all the vernier gaps). Also included in the program written for the Quaestor microscope was a routine for measuring linewidths for the purpose of adjusting exposure times more easily and accurately. It was observed, however, that different layer materials would show up with different contrast levels and field depths, therefore leading to wrong readings. Calibration curves were then prepared for the different cases (photoresist over SiO_2, photoresist over polysilicon, polysilicon over SiO_2, etc.) using a Vickers electronic microscope with automated linewidth measurement facilities and then fed into the Quaestor microscope. An example of such curves is seen in Figure 4.12. A typical automated linewidth measurement is shown in Figure 4.13.

The second photolithographic step revealed what was to be one of the most critical problems of this process: there was a significant difference amongst the registration offsets read from the different verniers within a four chip group. This difference reached, in some cases, 0.3 μm. The initial explanation for this problem was that of field distortion. Two of the verniers are located at the periphery of the field while the third is located at a more central position (Figure 4.14). If any kind of optical distortion is present, such as spherical aberration, the effect will be more or less evident depending on the distance to the optical axis. This would therefore explain any differences between the reading of the central vernier and the other two readings. Indeed it was noticed that the two peripheral verniers seemed to give similar readings which would differ quite significantly from the central vernier. However, such aberration would affect all layers

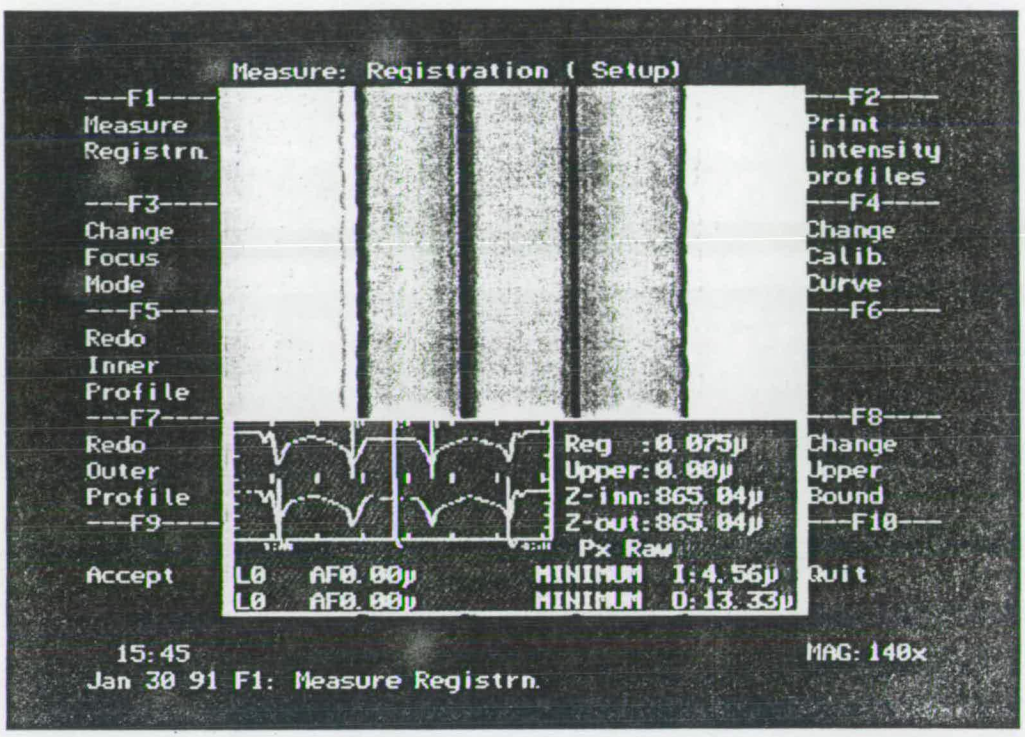


Figure 4.11: Example of an automated registration measurement using one of the vernier structures.

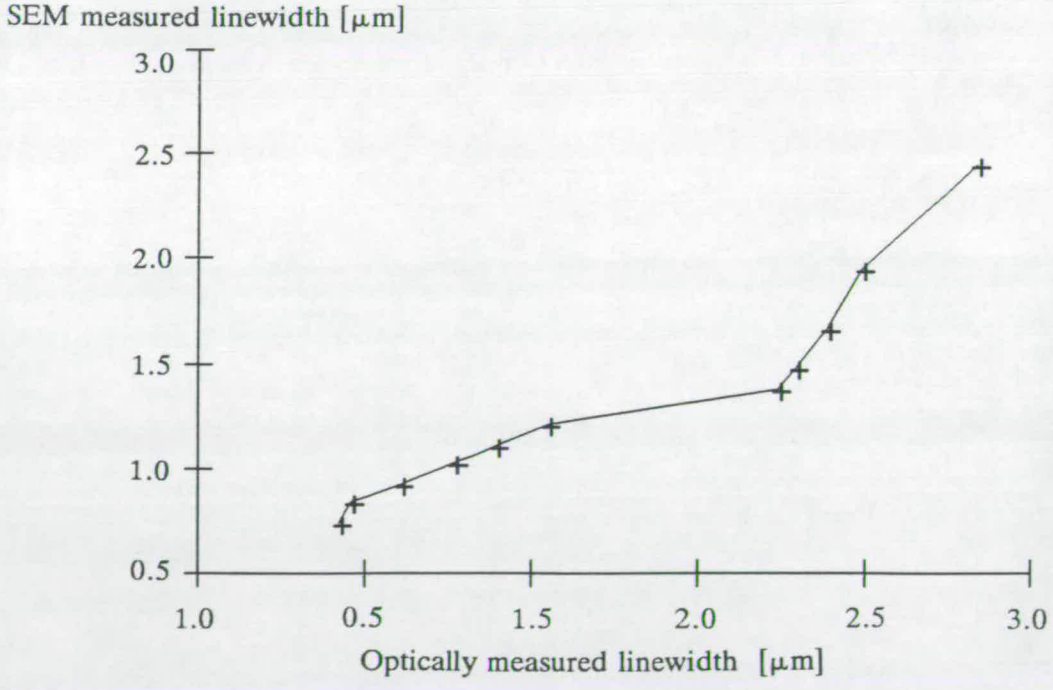


Figure 4.12: Calibration curve for photoresist over polysilicon.

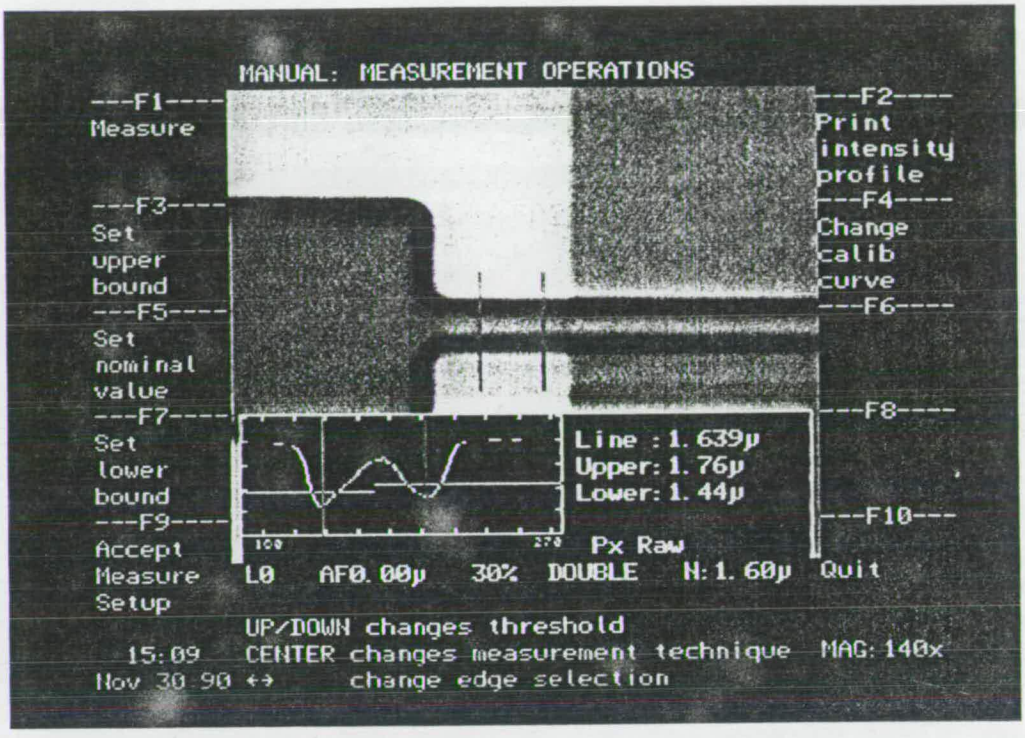


Figure 4.13: Example of an automated linewidth measurement taken at one of the transistor gates.

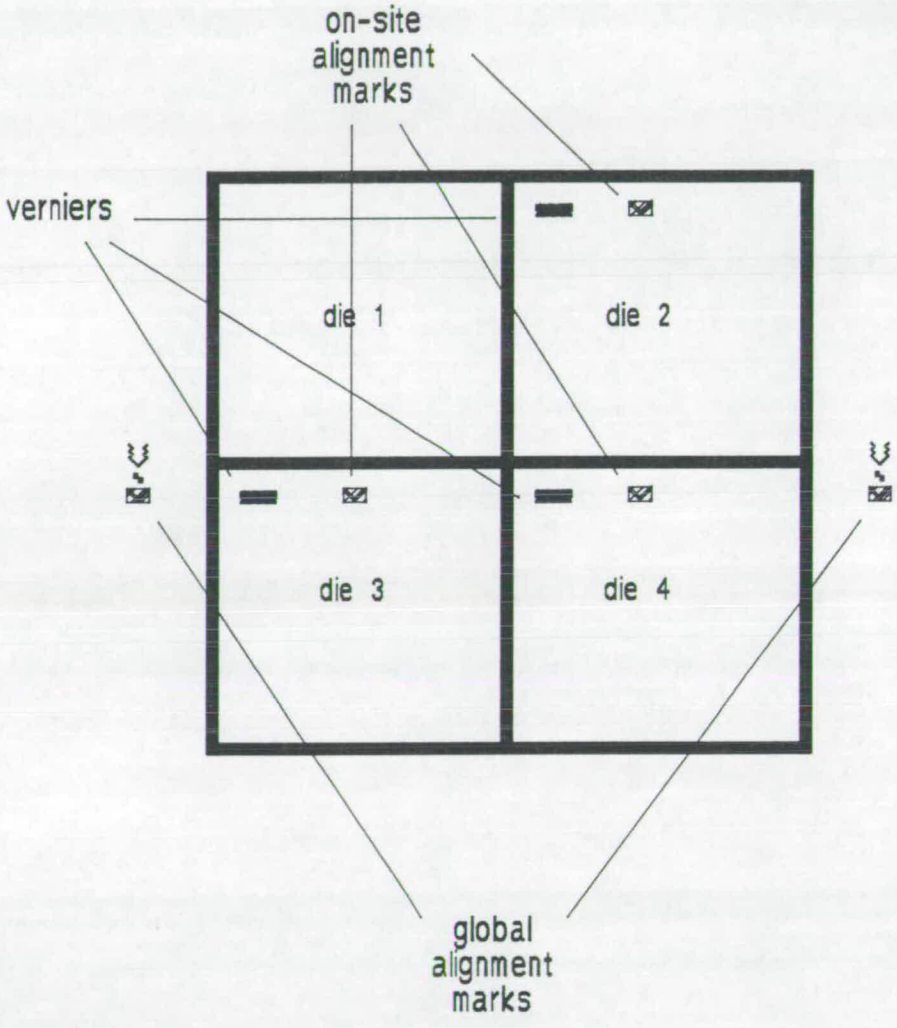


Figure 4.14: Positions of verniers and alignment marks on a chip.

equally and would also have been noticed in one of the periodic calibrations which are routinely undergone. It was therefore decided to try a few more experiments.

An Optimetrix wafer stepper was used in all photolithographic steps of this fabrication process. It allows a combination of different alignment techniques, of which a significant number were tried in an attempt to minimise registration problems. Figure 4.15 shows a simplified diagram of the wafer stepper. It has a main optical axis, consisting of the main illuminator, the reticle assembly, a beam splitter, the reduction lens and the stage assembly. The beam splitter allows the observation of the image projected onto the stage superimposed with the image of the feature already present on the stage, as represented in Figure 4.16. The optical system which allows this direct pattern matching observation is called the on-axis microscope. There is also an off-axis microscope which can be used to observe the matching of patterns printer on the wafer with a fixed target. The operation of the equipment usually begins with the insertion of a new reticle onto the reticle assembly. The reticle is aligned with the stage by matching its global alignment marks with the stage alignment marks (Figure 4.6) using the on-axis microscope. Note that the stage, not a wafer (there are no wafers on the stage at this point), is used for this alignment procedure. The alignment is done at first manually, using the the double-box structure and the split cross and then the internal computer does the fine alignment by matching the lower part of the "Christmas tree" structure with the corresponding framing pattern. The base line, i.e. the separation between the on-axis and the off-axis microscope is also calibrated. At this point the stage, optics and reticle are all aligned with each other. The next step is to load a wafer and align it to the rest of the system.

There are different methods to align a wafer with the stepper optics depending on how precise and fast the process is supposed to be. Wafers are printed with the same global alignment patterns depicted on the left hand side of Figure 4.6. The first step is always to align manually the wafer's global alignment marks (in this case the Christmas tree structure) with the frame projected by the fixed target. Once this is done, there are two main possibilities to follow: to use the same off-axis microscope for a further automatic alignment or to use the on-axis microscope for this fine adjustment. If the first is chosen, the same patterns used in the off-axis manual pre-alignment described above

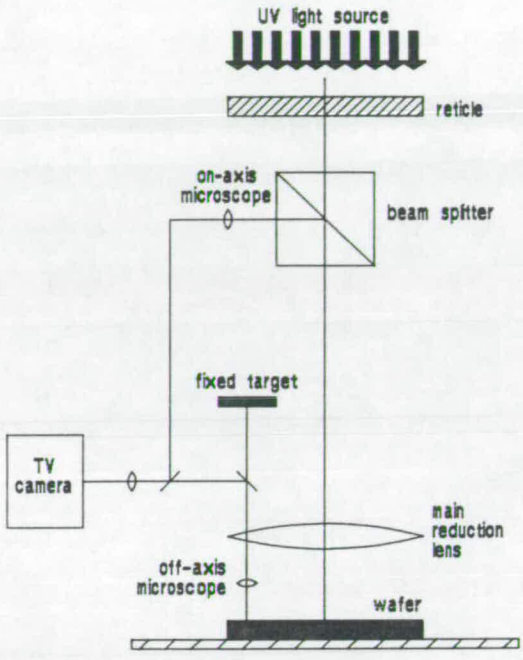


Figure 4.15: Simplified diagram of the wafer stepper (after [8]).

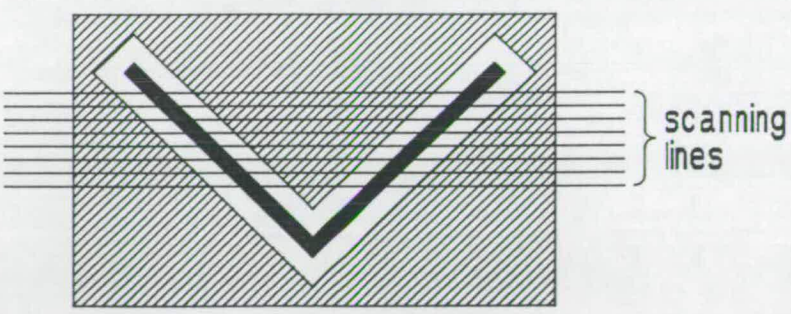


Figure 4.16: Overlayed image of an alignment mark. The scanning lines are used by the built-in computer to perform alignment. The tolerances for such operation as well as the number of scanning lines are among the parameters which can be adjusted to allow a better registration.

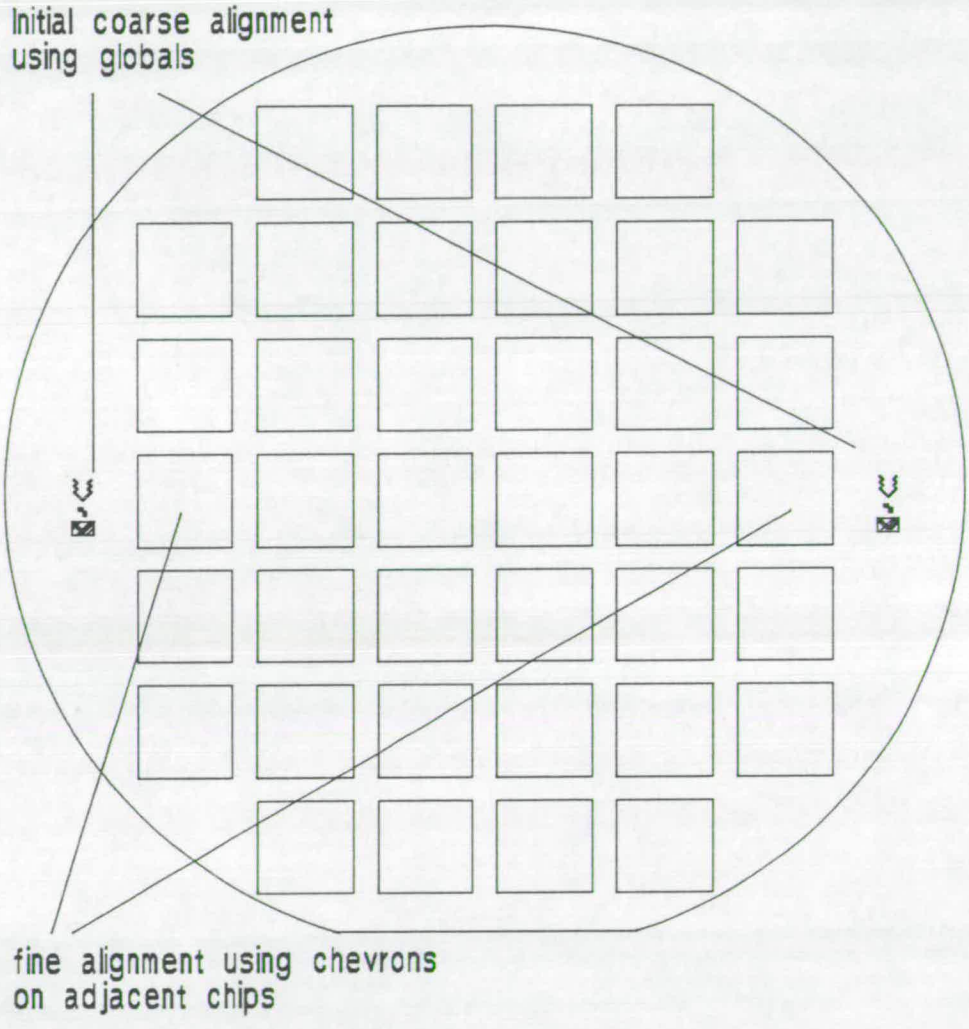


Figure 4.17: On-site global alignment technique. Once the coarse alignment is obtained using the global alignment marks, the chevrons in the chips adjacent to the globals are used for a fine adjust.

are used by the internal computer. If the second option is chosen, the alignment pattern contained in the chips immediately adjacent to the global alignment marks are matched against the pattern in the reticle (Figure 4.17). The advantage of this technique over the previous is obvious: the off-axis alignment is an indirect method (the wafer alignment marks are matched against the fixed target and the reticle is matched against the stage) which relies, amongst other things, on the fact that the base line calibration is absolutely accurate, whilst the on-axis method uses the direct alignment between reticle and wafer.

Once the alignment between wafer and reticle has been achieved, the wafer is ready for the die-by-die exposure (in the present case die should be understood as our four chip ensemble). This can be done in two different ways, as far as the alignment is concerned: a blind step, which consists in simply carrying out the exposure steps with no further alignment operations, and the die-by-die alignment, which includes one automated on-site alignment (identical to the process described above for the on-axis fine global alignment) preceding each exposure. Although this technique may sound ideal, it has one major limitation. The automated global alignment described before includes, in both cases (off-axis and on-axis) theta (rotation) adjustments as well as x and y adjustments. When the die-by-die alignment technique is chosen, however, any further necessary corrections only include x and y adjustments. It also has the disadvantage of slowing throughput considerably.

Apart from the choice of alignment techniques, there are many other variables that can be adjusted for optimum registration. In order to understand some of the most important, we must also understand how the internal computer does the automated alignment. Figure 4.16 shows the on-site alignment mark, superimposed on the matching pattern projected from the reticle. This chevron is scanned by video lines whose number can be controlled by the operator, as well as the tolerance required for the alignment. Also within the operator's control is the number of different attempts amongst which the best alignment will be chosen. Another important aspect to consider when the alignment is critical is the sharpness of the chevron's image. Due to the fact that there is photoresist on top of the on-site chevron, its image will show up not as sharply as it would be had the surface been perfectly clear. The effect of such coating may vary according to the layer in question, but generally speaking it is detrimental to a good alignment. There are

two different ways to circumvent this: delaying the alignment step or clearing the chevron areas beforehand. The first technique relies on the fact that the exposure exerts a "bleaching effect" upon the photoresist. After the exposure, all feature contours become much sharper and contrasts more noticeable. It is sufficient therefore to expose the chevron area for a time long enough for this effect to take place (usually around one second) prior to the alignment routine. Since this pre-exposure takes place through a slit which frames out the rest of the die, there is no prejudice in terms of overexposing it. The other technique consists of exposing the chevrons in the same way described above and developing it. This will remove all the photoresist on top of the alignment marks without affecting the rest of the die areas (positive photoresist is being used for all layers). Wafers are then reloaded into the stepper and the alignment/exposure process follows normally.

All the above techniques and variations were tried in the course of the fabrication process described here. The objective of such experiments was three-fold: to minimise all the individual misalignments, to minimise the variation in misalignments in a wafer scale and to minimise the difference amongst misalignments within a single die. The best alignment technique varied from layer to layer, so for each photolithographic step the alignment experiments had to be repeated. The automated global alignment using the on-axis microscope followed by blind stepping seemed to be marginally better than the others, although there were layers for which it seemed to be consistently bad. Exhaustive tests had therefore to be carried out at every photolithographic step. In some cases the misalignments seemed to be consistently repetitive, so offsets were introduced in order to compensate them. The main reason for concern, however, was that there were large differences amongst the three verniers of each group of four chips (or die) already mentioned before. Because of these local differences the special transistors for alignment assessment which were included in chip one were not used since any eventual misalignment measured through them would apply only to the region immediately surrounding the transistors.

After exploring most of the stepper capabilities for alignment optimisation, it was found that in some cases it was impossible to achieve a good agreement between vernier measurements within a die. Apart from the field aberration theory, the

distortions could also be explained by inferring that the reticles themselves have been distorted. This could have been caused by a bad repeatability of the pattern generator employed or by thermal dilations affecting the reticles themselves. However, two important precautions had been taken beforehand: all the reticles were fabricated in sequence (so as to find the pattern generator in roughly the same operating conditions) and kept together under the same environmental conditions. The most likely hypothesis, however, was raised in the course of the alignment experiments and mainly in the second process which will be discussed later in this chapter: rotation mismatch. Because no further rotation adjustments are made after the global alignment, any remaining rotational error will continue even if special die-by-die techniques are used. It was not until the second process that it was found that it is possible to improve this rotational alignment. Furthermore, the small number of verniers in a die made it particularly difficult to analyse such problems in more detail.

4.2.2.2. Analysis of the contact implant

One of the reasons why silicon is the predominant semiconductor for the fabrication of microelectronic devices lies on the fact that its thermally grown oxide is of very good quality (it is very uniform and exhibits very good electrical characteristics) and is easy to grow. However, in at least one circumstance this can represent a problem. When contact windows are etched to create contacts between silicon and the aluminium layer, the exposed silicon immediately starts to oxidise, creating a very thin layer of native oxide.[1] Aluminium is then deposited and a sintering step is run in order to reduce this layer and create ohmic contacts. While some spots of this thin oxide will eventually disappear faster, another detrimental factor also comes into play: the difference in solubilities of silicon in aluminium and of aluminium in silicon. Silicon has a solubility of 0.5% by weight in aluminium at 450°C (the average temperature at which sintering is done) and the solubility of aluminium in silicon is lower than 0.001% by weight,[2] which means that silicon will diffuse into aluminium, particularly at the sites which are being cleared of native oxide first (in the beginning of the sintering process there is no silicon dissolved in the aluminium layer, so the exposed silicon spots will find it very easy to diffuse into the aluminium). The diffused silicon leaves behind voids

which are filled with aluminium, thus creating spikes. The formation of these spikes has two negative consequences. First it leads to very high electric fields at the extremity of such spikes. This is not so serious in most of the cases because the junctions are usually doped heavily. Second it can eventually trespass the junction and cause a short circuit between the contact and the underlying silicon substrate. This is of utmost concern for the fabrication of small devices, not only because of their shallow junctions but also because of the presence of small contacts (smaller contacts tend to have more aluminium on top of them which makes it easier for silicon to diffuse into it).

This effect (sometimes referred to as *Kirkendall effect*),[2] can be minimised in different ways. One of the most common, which has been used in this work, is to use aluminium with a small percentage of silicon, usually between 2 and 3% by weight. Nevertheless the problem is not totally eliminated because of the high diffusivity of silicon in aluminium. The next most common approach to circumvent the problem is to make sure that the junctions are deeper than the spikes in the regions concerned. A contact implant is then done, usually using phosphorus in order to achieve the maximum possible junction depth beneath the contact region.

The present fabrication used shallow junctions (in the order of 0.25 μm deep). Furthermore, there were exposed active areas in the offset transistors, so there was a risk of those regions getting implanted. This is because the only material on top of the exposed active areas is the pyrolytic oxide used for passivation. Pyrolytic oxide offers much less protection against ion penetration than thermally grown oxide. For an implanting energy of 150 keV (which is not even high enough to produce the desired effect of a deeper junction) the required thickness of thermal oxide in order to stop the ions is 5000 Å,[3] which gives an idea of how ineffective the 7000 Å passivation layer would be against even higher energies. The other factor to be considered is the risk of enhancing spiking by generating more defects if too high an energy and dose are used. High energies lead to deep lattice damage and high doses cause the damage to spread over a range of depths. Although further high temperature steps are executed (at least one such step is necessary after the contact implantation in order to activate the implanted species) not all the damage is annealed. It is therefore important to avoid excessive doses and energies.

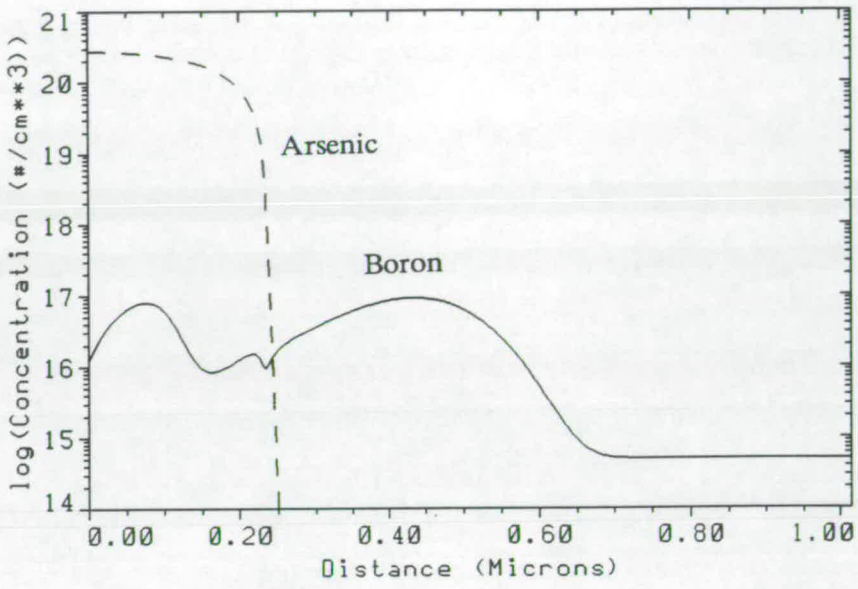


Figure 4.18: Simulated doping profiles (boron and arsenic) immediately preceding the phosphorus implant.

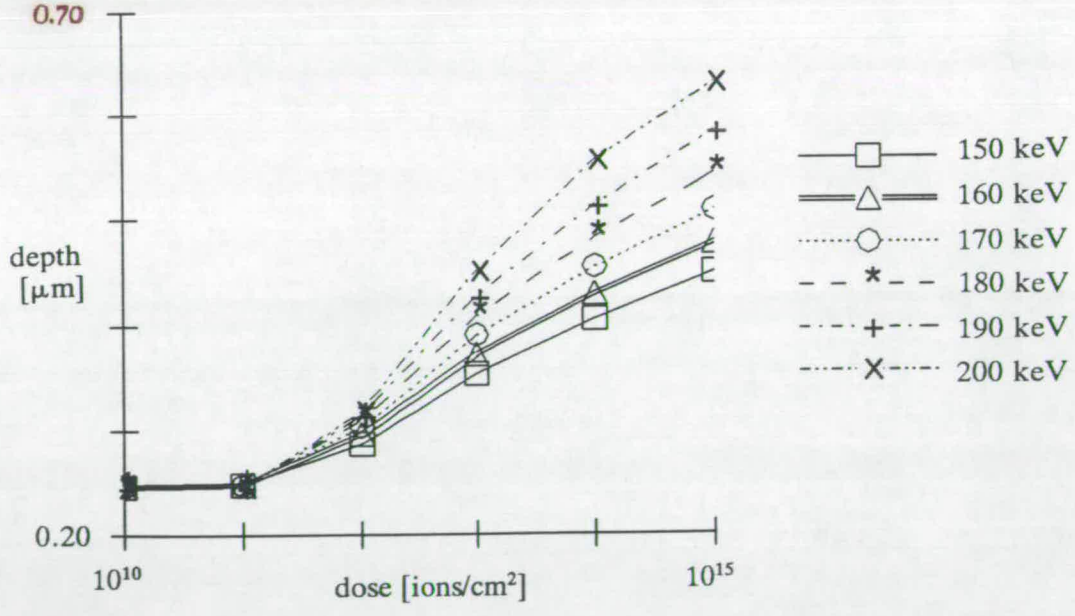


Figure 4.19: Junction depth versus phosphorus implant dose with implanting energy as a parameter.

One of the difficulties of designing the contact implants is the shape of the boron profile created earlier in the process. As can be seen in Figure 4.9, a double boron implant is done right after the gate oxide is created. The shallower implant has the purpose of adjusting the threshold voltage whilst the deeper implant minimises punchthrough. Apart from being deeper, this second implant also has a higher dose. Deeper junctions become therefore more difficult to make (as depicted in Figure 4.18), due to the higher energies and doses required. Figure 4.19 shows the junction depths which can result from implants ranging from 1×10^{10} to 1×10^{15} ions per square centimetre in dose and from 150 to 200 keV in energy. Once the phosphorus concentration surpasses arsenic at a dose level of about 1×10^{11} ions/cm^2, the junction depth increases sharply but starts to saturate for doses higher than 1×10^{13} ions/cm^2. This saturation is caused by the boron layer.

A 180 keV, 1×10^{13} ions/cm^2 implant was finally chosen. Due to the fact that this energy lies too close to the limits of the Lintott implanter used, doubly ionised phosphorus was selected as the implanting species so that an accelerating voltage of 90 keV could be used. The wafers were protected with high energy resistant photoresist (OFPR 800) so that the active areas of the offset transistors would not be affected (only the contact areas were left exposed). A 1.5 μm thick photoresist layer was used, which is enough to stop the ions (a 1.0 μm layer has a transmission of 0.0001% [3]).

As it will be further discussed in Chapter 5, it was detected during the preassessment of this batch that the self-aligned transistors presented problems of junction integrity. This was initially attributed to an insufficiently deep contact implant. However, this effect was not at all noticed in the offset gate transistors, not even after long term stress cycles. As it can be seen in Figure 4.8, the two types of transistors had almost the same processing conditions, except for one: the offset gate transistors had their source and drain implanted and were then covered with polysilicon, whilst the self-aligned transistors only had their source/drain implant after the polysilicon had been deposited and patterned. This means that the offset gate transistors underwent one extra high temperature step: the polysilicon deposition. Although this temperature is not so high compared to the other steps, it is high enough to promote annealing of the damage caused during the arsenic source/drain implant, particularly because of its long time.

Therefore it might have happened that the subsequent high temperature steps were not enough to anneal the implant-generated damage in the case of self-aligned transistors. If that had been the case, it was probably difficult for the phosphorus ions to get implanted deeper than the damaged region, thus leaving the device unprotected against spiking.

4.2.2.3. Polysilicon recrystallisation

In the present process polysilicon is deposited at 600°C. It is a well known fact [3] that this is a critical temperature; below it, polysilicon tends to assume an amorphous structure and above it it becomes granular. In case the amorphous form gets deposited, a subsequent high temperature step will turn it into a granular structure, with grain sizes tending to be larger than they would be if the initial structure were granular. This crystallisation process could have taken place during the reflow and densification of the pyrolytic oxide. However, an earlier crystallisation could also have taken place during the second arsenic implant (for the definition of source/drain of the self-aligned transistors). This hypothesis has been raised because of the spots observed over the active areas of several offset transistor samples (Figure 4.20). Those spots could be the consequence of the damage caused on the oxide by the formation of crystals at the polysilicon/SiO_2 interface. It is important to mention, however, that this effect has not affected device performance. This could be attested by the fact that not all wafers were affected by it, so samples with and without the spots could be compared.

4.3. Second fabrication

Three main factors led to the decision of starting a second fabrication batch: the difficulty in determining the LDD dimensions in photolithographically defined LDDs, the alignment distortions observed in the first process and the need for smaller transistors to complement the results obtained with the initial samples. The first factor implied a totally different approach to the LDD problem, whilst the second and third factors meant that the design also had to be modified.

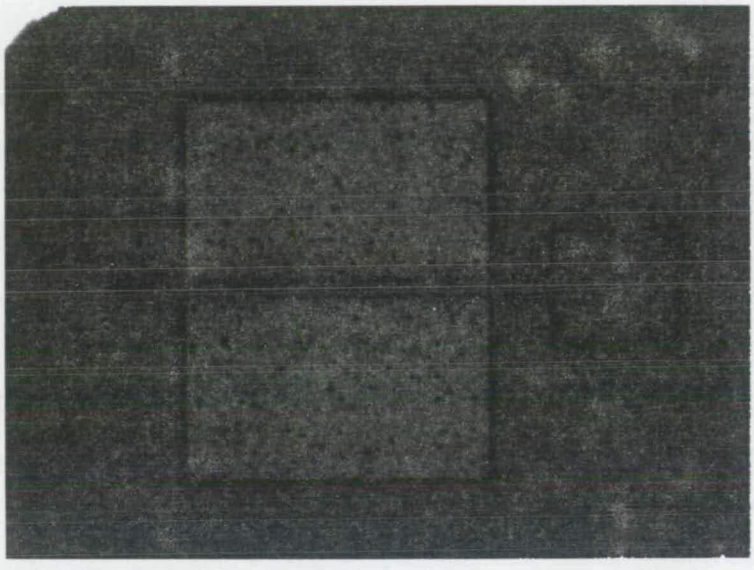


Figure 4.20: Spot effect observed in several offset-gate transistors after the second gate patterning.

4.3.1. Design considerations

Different hypotheses for the alignment problem observed in the first process were discussed earlier in this Chapter. Whatever the cause was, it is quite obvious that a smaller field area would be less affected. A 5 by 5 millimetre chip size (as opposed to the 10 by 10 millimetre size used previously) was then selected. Ten arrays of offset gate transistors were included, each array with a different channel length. Channel lengths ranged from 0.6 to 1.5 μm and gate offsets from -0.5 to +0.5 μm, all variations in 0.1 μm steps. An array of special structures for registration measurements was also included. This structure, depicted in Figure 4.21, was designed such that it would be easily recognisable by the Quaestor's pattern recognition system, therefore making the measurement faster and more reliable. In order to facilitate sample cross-sectioning, a "long transistor" was also incorporated. This transistor, although non-operational, crosses the entire chip so that almost any cross-section in the transverse direction would intercept it and show surface topography, impurity profiles, etc. A big capacitor for C-V measurements and the usual on-chip alignment marks (chevrons) were also included.

The mask set was constituted by the following layers:

Layer 1: active areas
Layer 2: source/drain definition
Layer 3: gate cut
Layer 4: contact windows
Layer 5: aluminium tracks
Layer 6: passivation windows

Figure 4.22 shows the alignment sequence for these layers.

4.3.2. Fabrication process

Once again the fabrication process was based on EMF's 1.5 μm nMOS standard process. As in the first batch, the process had to be modified for the fabrication of offset-gate transistors. This time, however, no self-aligned transistors were included, therefore simplifying the process and reducing the number of masks. This number was further reduced by the modification of the process with which the LDD regions are

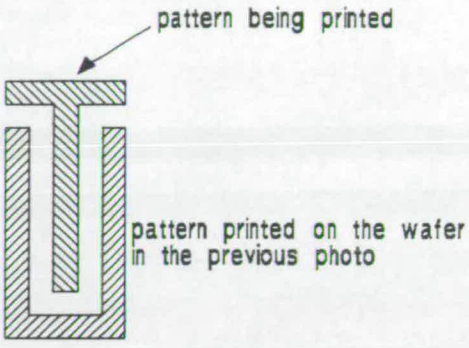


Figure 4.21: Pattern chosen for automated registration measurement in the second design. This pattern can not only be measured reliably but can also be easily recognised by the pattern recognition system of the Quaestor microscope.

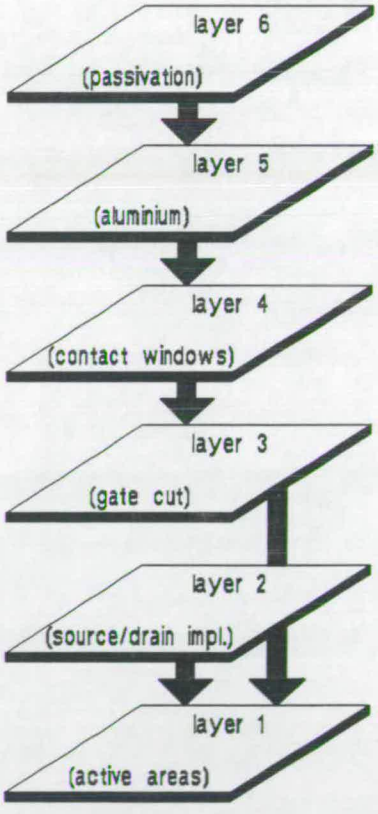
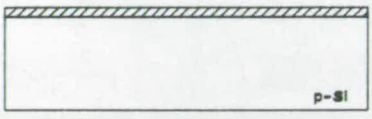


Figure 4.22: Alignment sequence for the second fabrication.

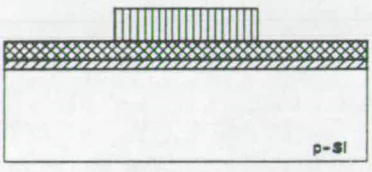
Initial oxidation (5 minutes O_2+HCl / 5 minutes H_2+O_2+HCl / 5 minutes O_2 at 950°C)
Target thickness = 400 Å



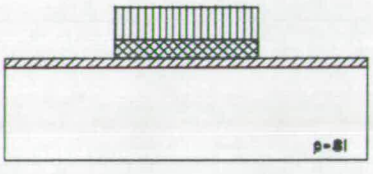
Silicon nitride deposition (20 minutes in SiH_2Cl_2+NH_3 at 200 mTorr, 850°C)
Target thickness = 1000 Å



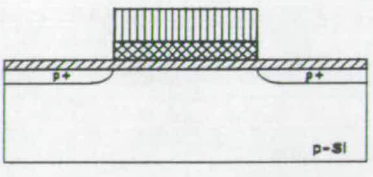
First photo (active areas)



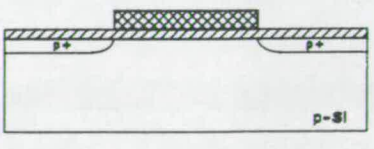
Si_3N_4 reactive ion etch - 3.5 minutes in CF_4:H_2 plasma



Boron implant (isolation implant) (dose=7×10^{12} ions/cm^2, energy = 70 keV)



Photoresist removal with fuming HNO_3

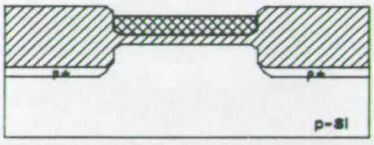


Pre-diffusion clean - 15 seconds in 10% HF

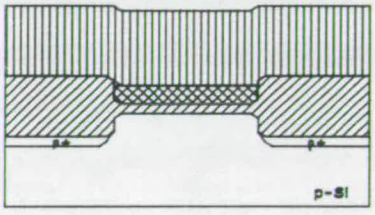
Karo's clean - 10 minutes in boiling 2:1 H_2SO_4:H_2O_2

Figure 4.23: Description of the second fabrication.

Field oxidation (5 minutes O_2+HCl / 30 minutes H_2+O_2+HCl / 3 hours H_2+O_2 / 5 minutes O_2 at 950°C)
Target thickness = 6000 Å

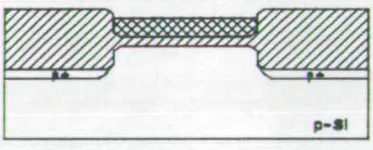


Photoresist coat (to protect the front of the wafers)



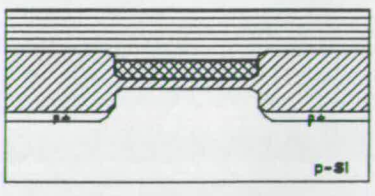
Back etch - Immerse in 4:1 NH_4F:HF until the backs dewet

Photoresist removal with fuming HNO_3



Oxide etch (to remove the top oxynitride) - 15 seconds in 4:1 NH_4F:HF

Polysilicon deposition (20 minutes in SiH_4 at 200 mTorr, 600 °C)
Target thickness = 2000 Å



Total oxidation of polysilicon (5 minutes O_2+HCl / 2.5 hours H_2+O_2+HCl / 5 minutes O_2 at 950°C)

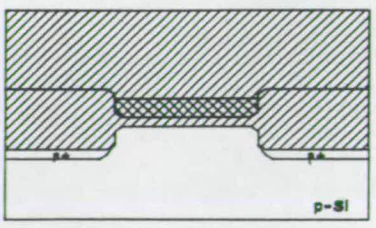
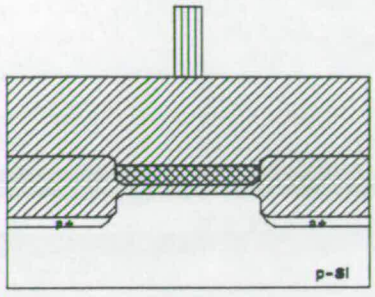
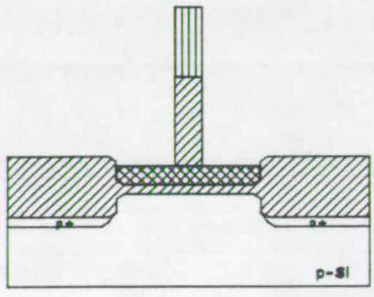


Figure 4.23: Description of the second fabrication (cont'd).

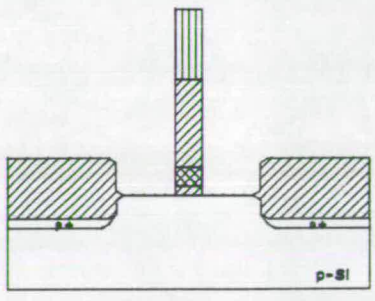
Second photo (barrier)



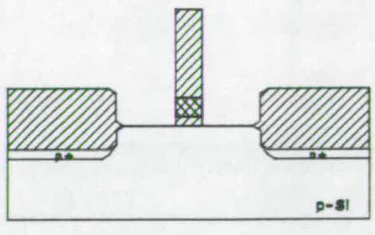
SiO_2 reactive ion etch - 40 minutes in CHF_3:He plasma



Si_3N_4 reactive ion etch - 3.5 minutes in CF_4:H_2 plasma



Photoresist removal



Amorphisation oxidation (5 minutes O_2+HCl / 2 minutes H_2+O_2+HCl / 5 minutes O_2 at 950°C)

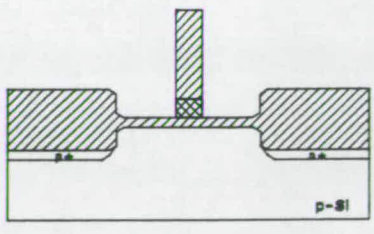
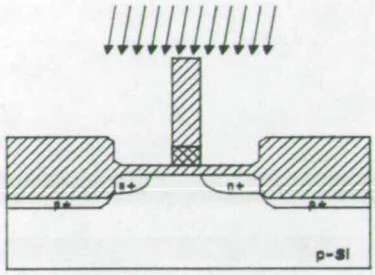
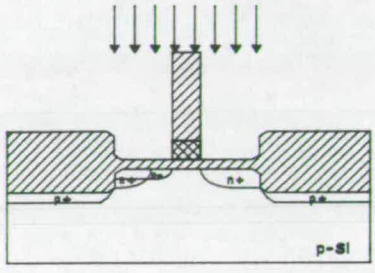


Figure 4.23: Description of the second fabrication (cont'd).

Arsenic implant (angled) (dose=7×10^{15} ions/cm^2, energy = 90 keV)

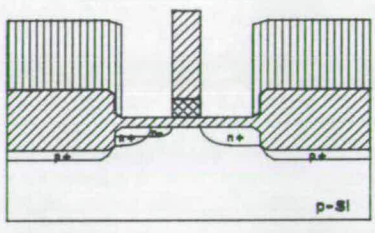


Arsenic implant (vertical) (dose=4×10^{13} ions/cm^2, energy = 50 keV)

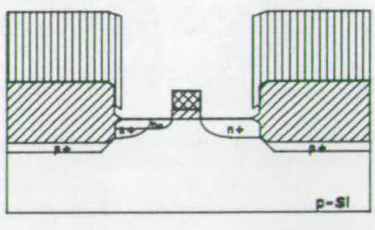


Implant anneal - 15 minutes N_2 at 950°C

Third photo - repeat photo with mask 1 using negative resist in order to protect field oxide.



Oxide etch (to remove the barrier oxide) 35 minutes in 4:1 NH_4F:HF



Photoresist removal using O_2 plasma

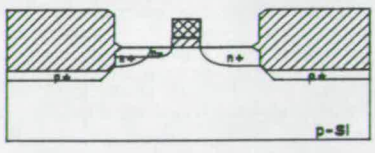
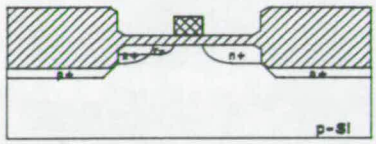
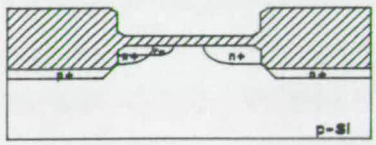


Figure 4.23: Description of the second fabrication (cont'd).

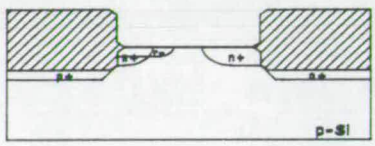
Protective oxidation (5 minutes O_2+HCl / 1 minute H_2+O_2+HCl / 5 minutes O_2 at 950°C)



Si_3N_4 wet etch - 30 minutes in H_4PO_4 at 165 °C

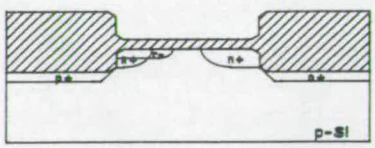


Oxide etch (to remove the protective and pad oxides) 3 minutes in 4:1 NH_4F:HF

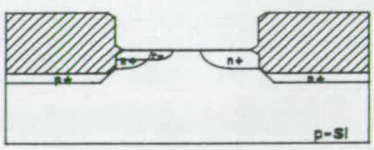


Karo's clean - 10 minutes in boiling 2:1 H_2SO_4:H_2O_2

Sacrificial oxidation (5 minutes O_2+HCl / 5 minutes H_2+O_2+HCl/5 minutes O_2 at 950°C)



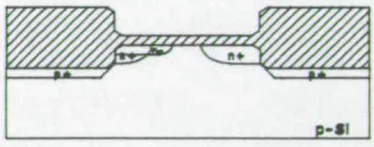
Oxide etch (to remove sacrificial oxide) - Immerse in 4:1 NH_4F:HF until the backs dewet



RCA clean - 10 minutes in 1:1:5 NH_3:H_2O_2:H_2O at 80°C / wash in DI water / 10 minutes in 1:1:5 HCl:H_2O_2:H_2O at 80°C / wash in DI water

Figure 4.23: Description of the second fabrication (cont'd).

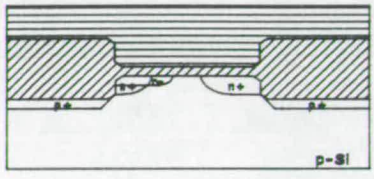
Gate oxide (5 minutes O_2+HCl / 2 minutes H_2+O_2+HCl / 5 minutes O_2 at 950°C)
Target thickness = 250 Å



Boron implant (threshold adjust and punchthrough control)
(dose=7×10^{11} ions/cm^2, energy = 25 keV
dose=2×10^{12} ions/cm^2, energy = 140 keV)

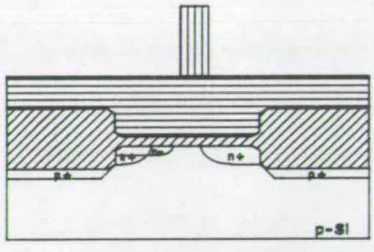
Karo's clean - 10 minutes in boiling 2:1 H_2SO_4:H_2O_2

Polysilicon deposition (40 minutes in SiH_4 at 200 mTorr, 600 °C)
Target thickness = 4000 Å

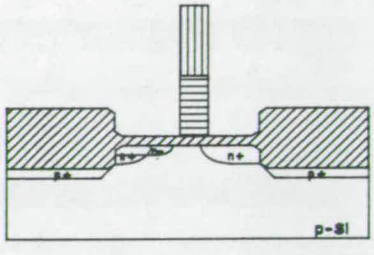


Arsenic implant (source/drain implant)
dose=7×10^{15} ions/cm^2, energy = 90 keV

Fourth photo (gate)



Polysilicon reactive ion etch - 10 minutes in CCl_4:N_2 plasma



Photoresist removal

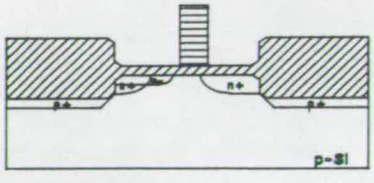


Figure 4.23: Description of the second fabrication (cont'd).

Pyrolitic deposition of SiO_2 (initial passivation)
SiH_4:O_2 reaction on a hotplate set to 430°C

Pyrolitic deposition of phosphorus-doped SiO_2
SiH_4:O_2 reaction on a hotplate set to 430°C
(with phosphine added)

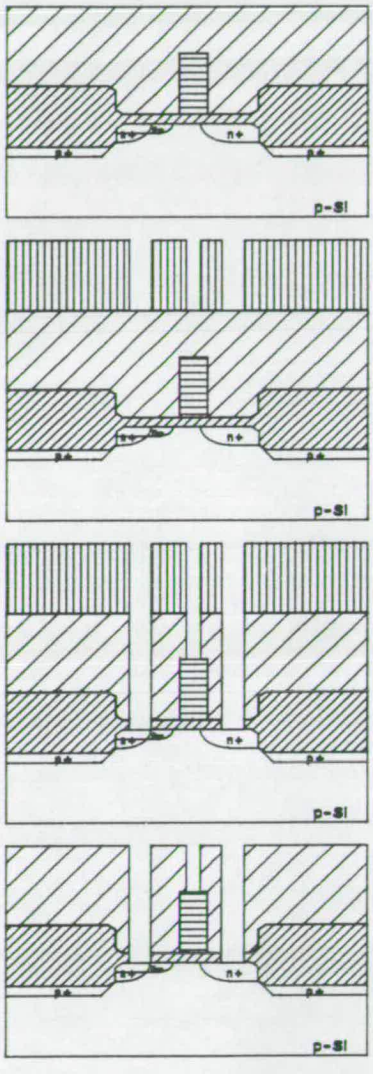
Reflow of the pyrolytic SiO_2 - 10 minutes in
an oven set to 1000°C with nitrogen flowing

Densification of the pyrolitic SiO_2 (5 minutes O_2 /
2 minutes H_2+O_2 / 5 minutes O_2 at 950°C)

Fifth photo (contact windows)

SiO_2 reactive ion etch - 40 minutes in CHF_3:He plasma

Photoresist removal

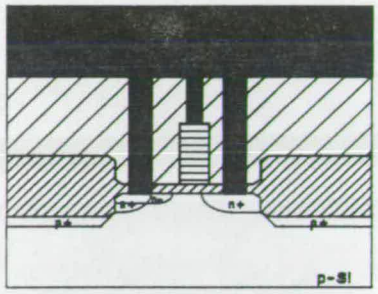


Phosphorus implant (spiking prevention)
dose=1×10^{13} ions/cm^2, accelerating potential = 90 kV
with doubly ionised species resulting in an energy of 180 keV

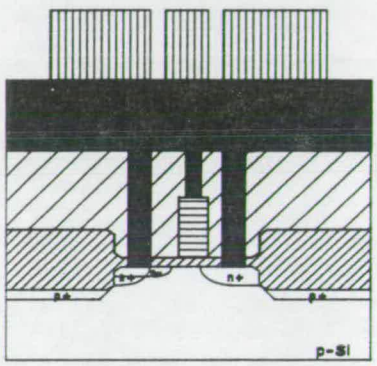
Pre-metallisation clean - 20 seconds in 25:1 NH_4F:HF

Figure 4.23: Description of the second fabrication (cont'd).

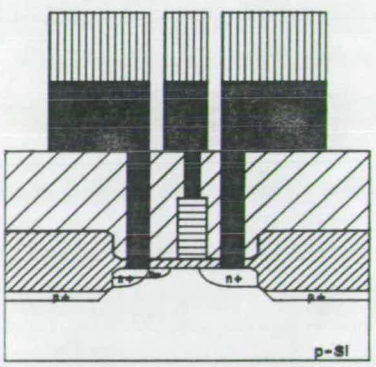
Aluminium sputtering - use of aluminium with 2% silicon
Target thickness = 1 μm



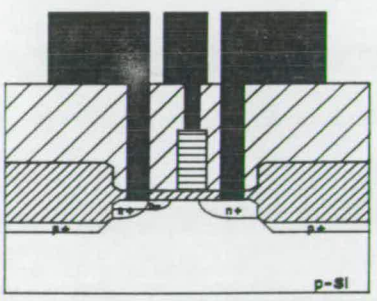
Sixth photo (metal patterning)



Metal etch - 5 minutes in CCl_4:N_2 plasma



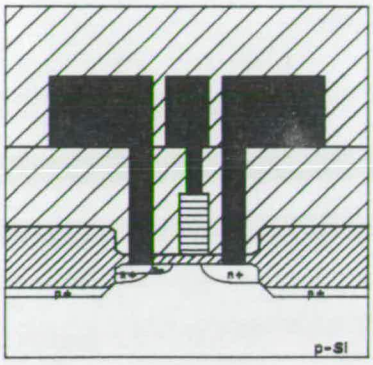
Photoresist removal



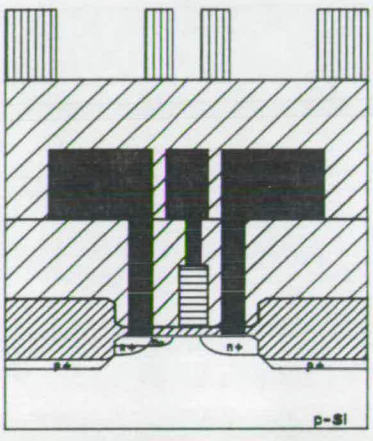
Contact sintering (5 minutes N_2 / 10 minutes H_2+N_2 /
5 minutes N_2 at 435°C)

Figure 4.23: Description of the second fabrication (cont'd).

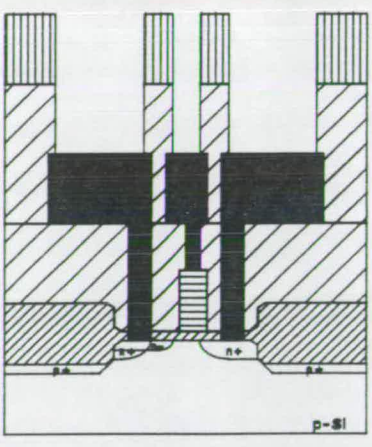
Pyrolitic deposition of SiO_2 (final passivation)
SiH_4:O_2 reaction on a hotplate set to 430°C



Seventh photo (pad window cut)



Pad window etch - Immerse in 4:1 NH_4F:HF until scribe lines dewet



Photoresist removal

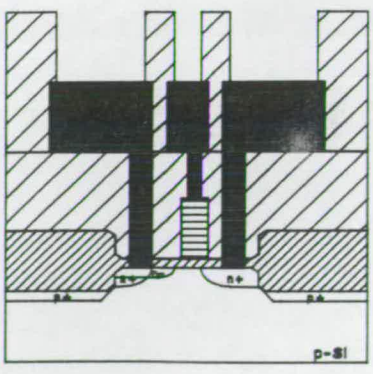


Figure 4.23: Description of the second fabrication (cont'd).

defined. The process modification, on the other hand, posed serious difficulties. In the end, the process had to be discontinued because of the uncertainty of the the final topography obtained. Figure 4.23 summarises the whole process.

4.3.2.1. Alignment control

Considering the experience acquired from the first fabrication, it was decided to concentrate the efforts in the improvement of the theta (rotation) alignment procedure. As was discussed before, no further theta alignment takes place after the initial global alignment, so blind steps were used throughout all the experiments and real runs of this process.

A series of wafers was tested initially using the unmodified alignment program in order to map the resulting error distribution. The concentration of larger errors on one side of the chip rather than at the periphery or in the centre clearly indicates that there is indeed a rotational error. Furthermore, this error was not consistent throughout the experimental batch as one might expect from a permanent optical distortion, thus suggesting that the tolerance in the alignment parameters was not good enough.

The global alignment parameters were then changed. Under normal conditions the alignment tolerance is set to 0.06 μm. If this tolerance is not achieved in the first attempt a retry is given. The parameters were then readjusted so that the tolerance was reduced to 0.03 μm and up to eight re-alignments were allowed. The stepper was also set to the expert mode, in which it improves the alignment routine by adjusting the width of the windows within which the alignment is detected, by determining the signal-to-noise ratio and adjusting the image sharpness. Additionally, the number of scanning lines across the chevron was increased from the usual 16 lines to 24. The results were further improved by adding offsets to account for any possible systematic errors.

4.3.2.2. Narrow line photolithography

There are two main factors which limit the capacity of producing very fine lines: the stepper's optics and the resist contrast. No matter how accurate the stepper system is, there is a physical limitation to the minimum feature size that can be projected. The

minimum feature size is limited by the wavelength of the light source; any attempt to go beyond it will be prone to diffraction. In the course of both the first and the second fabrication runs, focus and exposure experiments were carried out not only to obtain the right linewidths (according to the original designs) but also to optimise the operating conditions, therefore getting closer to the theoretical limits. Although the resist contrast does not affect the minimum feature size, it can lead to features larger than the initially projected image once the pattern is transferred to the underlying layer, e.g. by etching. When a pattern is projected on photoresist, the incoming monochromatic light suffers interference with the rays reflected by the underlying surface (beneath the photoresist). This will result in the formation of standing waves which in turn will create a tapered profile once the photoresist gets developed. This profile will have larger dimensions at the interface between the photoresist and the underlying material than at the surface (Figure 4.24). Tapering is further aggravated by diffraction in the case of small geometries.

Photoresist tapering can be greatly reduced by the optimisation of the developing process. Positive photoresists consist of a mixture of three main ingredients: a soluble novolac resin, a photo-sensitive dissolution inhibitor (an orthoquinone diazide) and an organic solvent (e.g. xylene). The resin is what constitutes the "body" of the photoresist. The organic solvent simply makes it fluid enough to be easily dispensed. The inhibitor has the property of preventing the resin from being dissolved. If it gets exposed to light, it loses its capacity of protecting the resin. After the the photoresist is dispensed on the wafer, it is cured in an oven or on a hotplate. The curing removes the solvent, which has now already served its purpose. The photoresist then shrinks slightly because the solvent constitutes a considerable percentage of its volume. It is obvious that such a curing process starts from the photoresist's top surface (where the solvent vapours can be dispersed in the air) towards its bulk. However, if the temperature is not high enough, the curing process is not completed and some solvent is left in the bulk of the photoresist (Figure 4.25.a). When the photoresist is exposed, the usual tapered profile will be present (Figure 4.25.b). If a second baking process takes place at this point, the rest of the solvent will evaporate, thus causing the bulk to shrink, consequently making the profile steeper (Figure 4.25.c). The other explanation for the post-exposure bake effect [4]

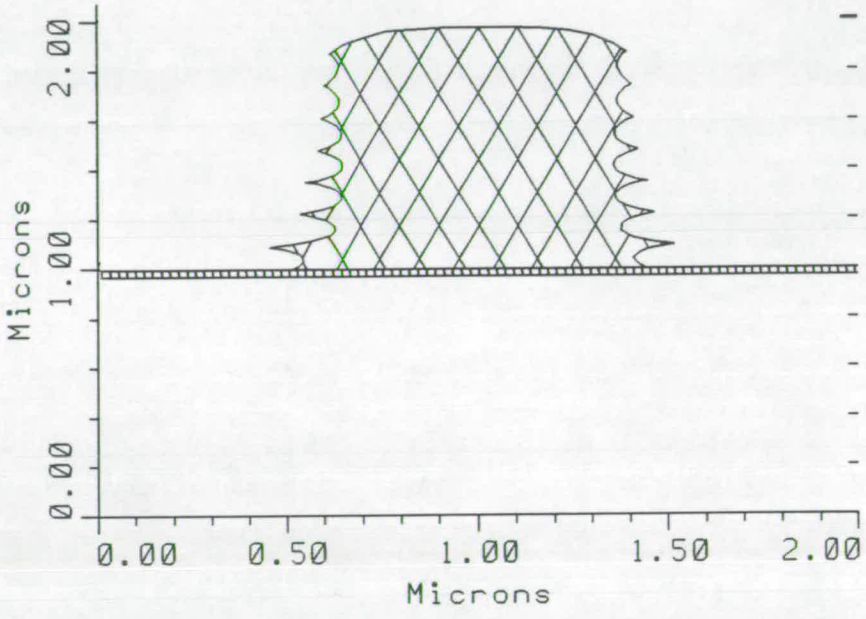


Figure 4.24: Simulation of the photoresist tapering effect.

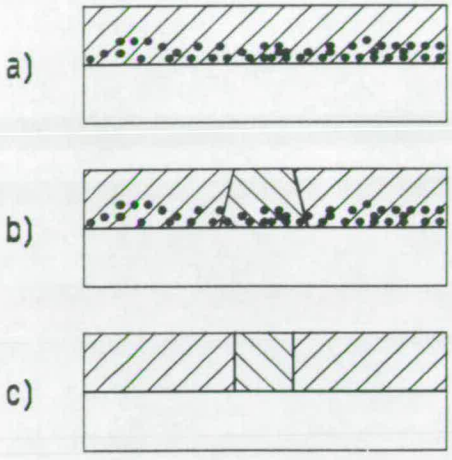


Figure 4.25: The solvent residue hypothesis for tapering reduction. a) A partially baked wafer has some solvent residue on its bulk; b) the pattern is printed and the usual tapered profile results; c) once the residual solvent is removed by further baking, the base of the pattern shrinks, therefore reducing the tapering.

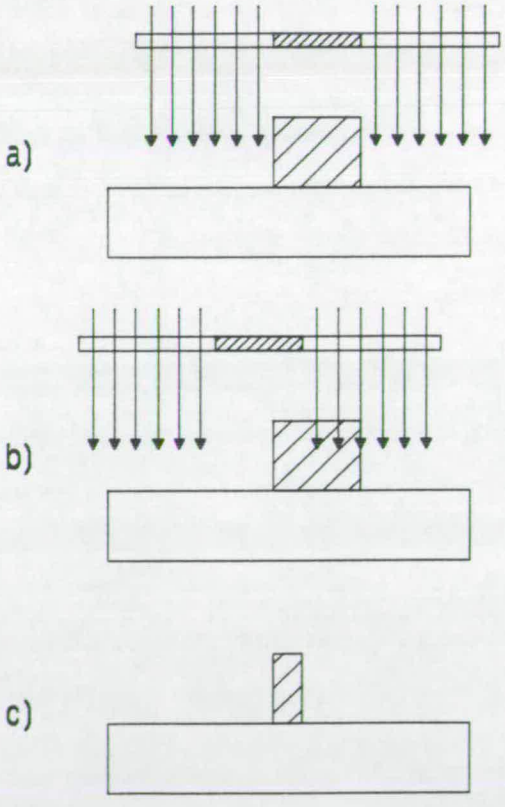


Figure 4.26: Double exposure technique. a) A wafer is exposed normally; b) a second exposure takes place, this time with a small mask offset; c) the intersection of the two projected images is kept, resulting in a much finer pattern.

is that when the standing waves are formed, there is no dissociation of the orthoquinone diazide in certain regions of the pattern edges, leading to the tapering effect. When the post-exposure bake is used, there is a diffusion of the dissociated orthoquinone diazide from the areas reached by light towards those zones which had been darkened by interference, therefore making the edges more uniform. In the present process, the usual pre-exposure bake temperature was reduced from 105°C to 75°C and a post-exposure bake at 120°C was introduced.

The optical constraint (i.e. the wavelength limitation) which prevents us from obtaining finer lines can be circumvented to a limited extent by taking advantage of the fact that positive resist is being used. If the photoresist is re-exposed with a small offset after the first exposure, the resulting patterns are the intersections between the shadowed regions left during each exposure (Figure 4.26). The size of such intersections will be limited only by the capacity of producing very small offsets and the amount of tapering in the photoresist. Care must be taken not to disturb the alignment between exposures, for it would be virtually impossible to reproduce the same alignment conditions twice. In the course of this fabrication process, fixed 0.5 μm offsets were used in order to reduce the 1.5 μm lines (the longer channel length in this design) to 1.0 μm, therefore also generating all the possible submicron transistor arrays up to the limit imposed by the tapering effect. It was absolutely mandatory to use blind steps so that the initial (global) alignment remained unchanged.

Another detrimental effect which had to be avoided is the flow of photoresist during the hard-bake process which follows the developing step. Photoresist usually has to be hard-baked before it undergoes processes such as ion implantation or plasma etching. Hard-baking is necessary in order to strengthen the resin. Unlike negative resists, positive photoresists do not undergo cross-linking between adjacent molecules, therefore resulting in a weaker resin. For this reason a high temperature is needed in order to promote this bonding. If such precautions are not taken, there is a risk that cross-linking will occur in an uncontrolled manner. In the case of plasma etching, for example, the process takes place only at the surface; since cross-linking leads to resin densification, the top surface tends to wrinkle and result in an uneven thickness. This effect is known as *resist reticulation*. The problem can be even more serious in the case of ion

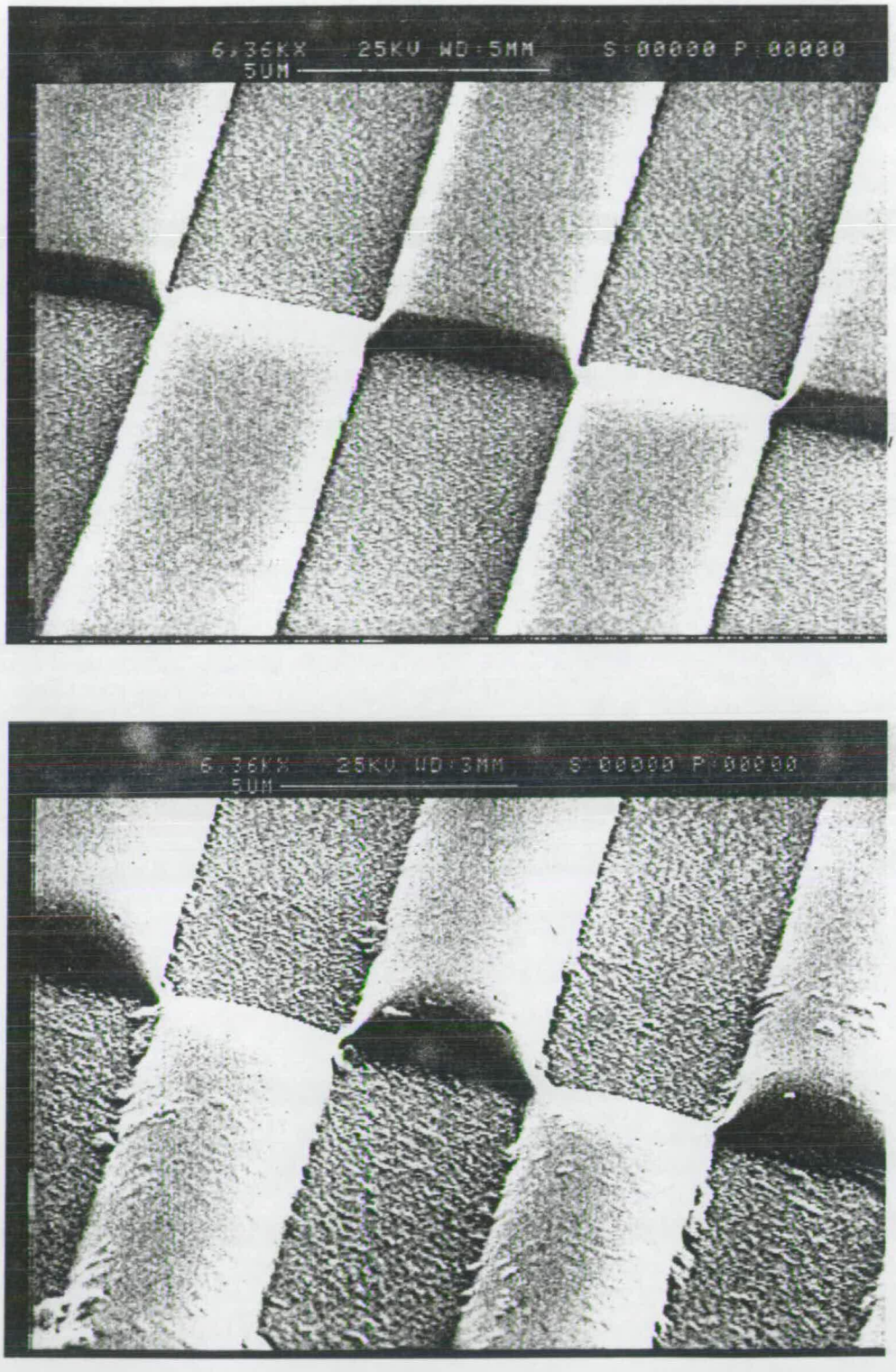


Figure 4.27: SEM picture of a resolution pattern printed with high energy resistant photoresist (OFPR 800) before (above) and after (below) undergoing a hardbake treatment.

implantation, where the photoresist is vapourised by the incoming ion beam. However, when a standard hard-bake process is used, the resin tends to flow and flatten the topography. Figure 4.27 shows the result of such a hard-bake on a previously steep pattern. Not only the feature dimensions become larger but also the photoresist thickness is decreased. This can be avoided to a certain extent by increasing gradually the baking temperature. Figure 4.28.a depicts the temperature profile used in the standard hard-bake process. For this fabrication a more gradual profile was designed, which is shown in Figure 4.28.b. With the modified fabrication process, fine lines such as the one shown in Figure 4.29 were obtained.

4.3.2.3. Implant barrier fabrication

Following the difficulties found in determining the amount of LDD overlapping in the photolithographically produced LDD transistors of the first fabrication process, it was decided that a different process had to be adopted for their fabrication. The main requirement for such process is that it had to provide an accurate way to determine the exact resulting LDD overlap during the fabrication step and so remove the need for a further LDD calibration. The possibility of adopting the standard spacer process had already been discharded during the first fabrication because it does not allow the use of systematic increments in the LDD dimensions (a single LDD overlap would have to be used for the whole process).

The solution chosen was to use an angled ion implantation to define the LDD overlaps. The technique consists of creating a wall to act as an implant barrier (Figure 4.30.a) and implanting with an angle different from zero with respect to the sides of this wall (Figure 4.30.b) †. The wall creates a shadow at the side opposite to the incidence of the beam; the length of this shadow can be adjusted by varying either the wall height or the angle of incidence. A second implant, this time perpendicular to the wafer plane, is used to create the LDD regions where shadows were left during the first implant (Figure 4.30.c). The process can be selectively repeated across a wafer by masking off different

† The usual 7° angle normally used in ion implants in order to avoid channeling is not being used in the present case and therefore is not being accounted for. It is being assumed that all channeling is being properly prevented by the use of oxide layers on top of the implanted areas.

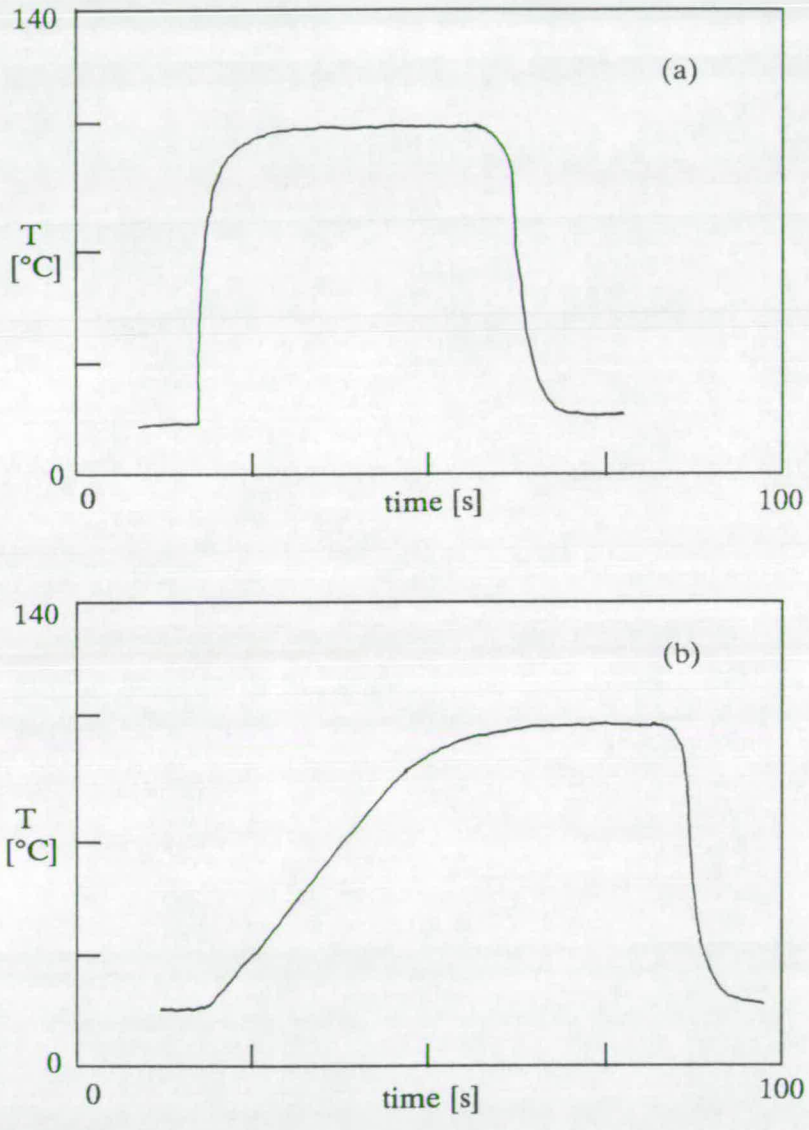


Figure 4.28: Temperature profiles used for hardbake. a) Standard process; b) modified (ramped) program for this fabrication.

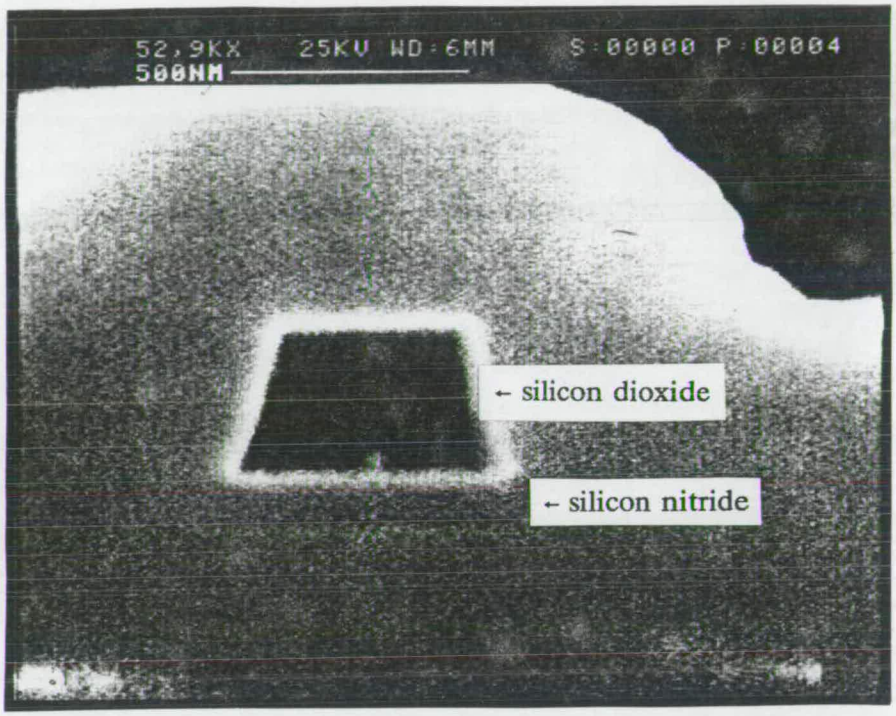


Figure 4.29: SEM picture of a 0.5 μm oxide line patterned with the double exposure technique and with photoresist hardbaked using the ramped program. Before being cleaved, the sample was coated with polysilicon and etched in HF in order to delineate the SiO_2.

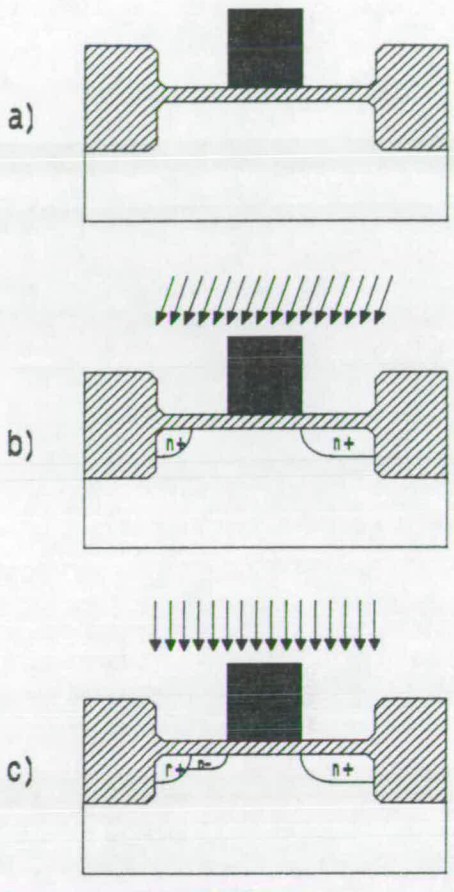


Figure 4.30: Priciples of the angled implant fabrication. a) A barrier is defined in the active area; b) a high dose, high energy implant is used in angle, so that the shadow casted by the barrier defines a space between it and the edge of the junction; c) a low dose, low energy implant is then performed vertically, defining the LDD region.

regions and repeating the angled implant using different angles, therefore producing different LDD overlaps in the same wafer.

The choice of the wall material was particularly difficult due to the characteristics required. Table 4.1 summarises the pros and cons of the possible candidates. Apart from having to exhibit good masking properties against the ion beam, the wall also had to have a good uniformity both in height across a wafer and in sidewall delineation. It also had to be easily removed once the implants were performed. Although different materials meet some of these requirements individually, one can hardly find a single material that meets them all at the same time.

Barrier material	Deposition	Masking thickness for a 0.0001% transmission (μm) [2]	Removal	Remarks
Photoresist	Easy	0.3	Easy	Flows when submitted to medium temperatures
Aluminium	Easy	0.15	Easy	Rough edges and surface
Silicon nitride	Limited to 300 Å	0.1	Easy	
Polysilicon	Easy	0.18	Difficult	Etch attacks bulk silicon
Silicon dioxide	Poor uniformity	0.15	Easy	Can be originated from Si or polysilicon oxidation (more uniform)

Table 4.1: Properties of different barrier materials

Photoresist is certainly one of the easiest materials to deposit and to remove. However, it does not stop ions very efficiently, which is particularly detrimental for angled implants. When the implant is vertical, the whole thickness of the masking material acts as a barrier for the incoming ions. However, if an angled implant is being used, the thickness of the barrier as seen from the point of view of the ions will not be uniform and the barrier will be easily penetrated at its thinner edges. Since the exact profile of the photoresist barrier will vary due to effects such as tapering, the exact size of the projected "shadow" (i.e. the resulting LDD overlap) will remain unknown. Additionally, photoresist tends to flow when submitted to hard-bake temperatures, as discussed earlier in this Chapter. Although this can be minimised, the originally sharp edges of the top surface can get rounded, therefore making the final result even less predictable. Photoresist can also be eroded at these weaker regions by the ion stream, again resulting in

undetermined shadow lengths.

Aluminium is certainly one of the best possible materials for stopping ions. It is also easy to deposit and strip. However, the character of its structure poses serious difficulties for its patterning. Its high reflectivity increases the interference problems during photolithography, leading to very poor feature profiles which severely limits the ability to produce narrow lines. Although this effect can be at least partially circumvented by the use of photoresist to which a dark dye has been added, there is still another major problem: the rough surface and edges produced during dry etch. This is caused by the presence of a native aluminium oxide which, if not totally uniform, will cause certain regions of the aluminium layer to etch faster than others. Different techniques are used to avoid such effect, like the use of thin aluminium layers (thin aluminium can be etched faster, thus reducing the chances for irregularities to form) and the use of pure aluminium (the presence of silicon can also lead to irregularities), but the final result will still be much worse than would be obtained with any other material.

Silicon nitride offers the best stopping power amongst the most commonly available materials in a standard MOS fabrication process. It is easily deposited by chemical vapour deposition (CVD) but can only be deposited up to thicknesses in the order of 2000 Å, beyond which it is likely to crack due to its high tensile stress.[5] It can be stripped using refluxed orthophosphoric acid (H_3PO_4), i.e. with the water contents being recycled, at temperatures in the order of 160 to 180°C.[6] It is important to keep a certain amount of water in solution because there is a dramatic decrease in selectivity between silicon nitride and silicon or silicon dioxide when other acids such as metaphosphoric acid (HPO_3) are present. The mechanism involved is relatively slow and leads to etch rates of the order of 100 Å/min against 10 Å/min for SiO_2 and 3 Å/min for Si. However, it was observed that implanted silicon has an etch rate considerably greater, which can present serious difficulties. The introduction of defects during arsenic implantation through SiO_2 has been discussed in the literature, but the origin of such defects is still unclear.[7]

Polysilicon is easy to grow and pattern, but poses a serious problem: anything that will etch it will also attack the substrate. It can only be safely protected by silicon nitride, which has its own problems in terms of layer removal. It can be protected by an

underlying pad oxide, but even this oxide layer has its own limitations, considering that it can also be attacked by whatever process is used to remove polysilicon and that it cannot be made too thick since the projected range of arsenic at the energies used in the present process start from about 400 Å.

Silicon dioxide can be deposited by different methods. At the time of the fabrication described here, the only available method in the EMF was the pyrolytic oxidation of silane. The reaction, namely

$$SiH_4 + O_2 \rightarrow SiO_2 + H_2 \qquad (4.1)$$

takes place at temperatures of the order of 400°C. The resulting uniformity depends on several variables and is usually very poor. Additionally, pyrolytic oxides do not have the same stopping power as thermally grown SiO_2. Thermally grown SiO_2, on the other hand, is usually very uniform and is extremely effective as an implant mask. Like pyrolytic oxide, it can be easily removed using diluted hydrofluoric acid (HF) which has a very good selectivity with respect to silicon. However, caution must be observed with both silicon and silicon dioxide (regardless of the growth/deposition method) after ion implantation has been carried out. Both materials can have their etch rates drastically increased if no annealing steps are performed. Even after annealing, an etch rate higher than normal should be expected.

The experiments for the fabrication of an implant wall started with an attempt to produce fine lines using aluminium. Different aluminium thicknesses, focus offsets, exposures and types of photoresist (with added dyes) were used without success. One of the best lines obtained is shown in Figure 4.31. It shows very clearly both the edge and surface roughness. The main reasons for concern in using such material for the real process were the impossibility of making very fine lines (therefore the impossibility of fabricating short channel transistors) and the irregular top edge which would inevitably lead to a irregular LDD edge (the projected shadow would be irregular).

The limited nitride thickness, together with the ease in growing thick layers of polysilicon led to the attempt of using a multi-layered structure for the implant barrier. Since polysilicon can be grown very thick, it would undoubtedly be the choice for the bulk of the barrier. However, in order to make it easy to be removed (considering the

impossibility of removing polysilicon without damaging the underlying SiO_2 and silicon substrate) and to increase the masking power, it was decided to include a layer of silicon nitride between the polysilicon layer and the pad oxide. Similarly, a layer of silicon nitride was included on top of the polysilicon layer to increase the stop efficiency. The resulting barrier therefore behaves as a nitride wall which can be built with any thickness, since the intermediate polysilicon layer can be as thick as necessary. The first problem to arise was that there is a poor selectivity between Si_3N_4 and polysilicon for both fluorinated and chlorinated gasses. It means that it is not possible to use photoresist to transfer a pattern to the top nitride layer and use it as a mask for the subsequent polysilicon etch. A new polysilicon layer was then added on top of the second nitride layer and oxidised in order to use the oxide as the masking material for the subsequent etches. The patterns were then transferred to the oxide using a CHF_3:He plasma. In etching the underlying layers, it was found that the top polysilicon, top nitride and bottom polysilicon etch without problems. When the bottom polysilicon/nitride interface is reached, however, the process does not go any further and the top layers start to overetch. It is believed that polymer formation takes place at that interface, therefore making it difficult to etch with the attempted plasma reactions. The possibility of using polysilicon as a wall filler was therefore discarded.

The other possible material for filling the space between two nitride layers and therefore form the bulk of the barrier was silicon dioxide. As mentioned before, pyrolytic dioxide has the particular disadvantage of being very irregular in thickness for this particular application. Given the fact that the only major disadvantage of polysilicon as a filler was the difficulty in etching, it was decided to adopt an intermediate solution: deposit polysilicon as in the previous experiment and oxidise it completely, therefore obtaining a very uniform layer of SiO_2. Considering that silicon dioxide is a very good masking material, it was also decided that no top nitride layer was to be used.

Experiments had to be carried out in order to determine the oxidation rate of polysilicon and make sure that the oxidation would be long enough as to achieve total transformation of polysilicon into SiO_2. A 900 Å pad oxide was grown on a few test wafers on which polysilicon was then deposited. A series of short oxidations was then performed, each time being followed by a photolithographic step in order to expose part

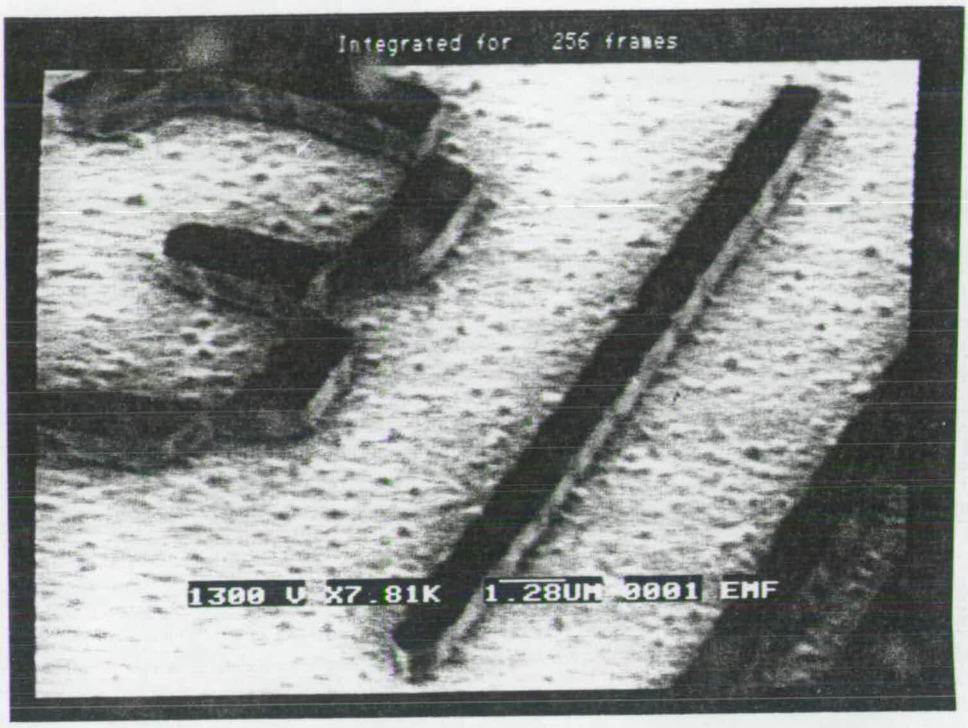


Figure 4.31: SEM picture of an aluminium line showing rough surface and edges.

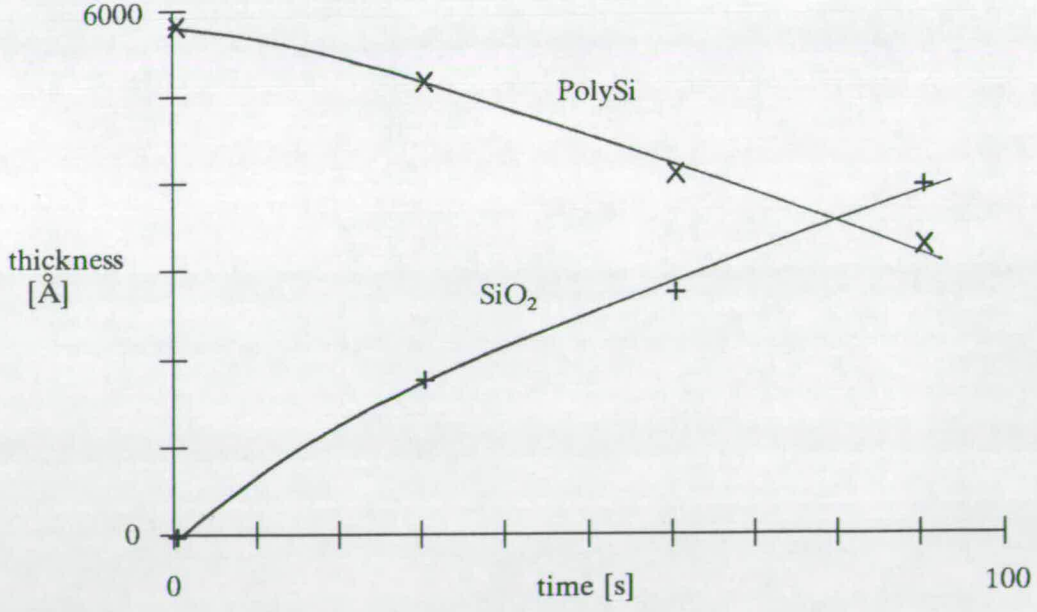


Figure 4.32: Oxide thickness and remaining polysilicon thickness versus oxidation time in the polysilicon conversion experiment.

of the superficial oxide. This exposed region was then etched and the photoresist was stripped off. The thickness of the remaining polysilicon was measured using a spectrophotometer and the oxide thickness was evaluated by measuring the height of the step constituted by the etched region and the adjacent oxide. The oxide thickness measurement was confirmed by measuring the time necessary to etch it. Three different etching solutions were used in the course of the experiments: HF 10% volume by volume, 4:1 buffered HF (NH_4F : HF) and 25:1 buffered HF, with etch rates of 100 Å/min, 1000 Å/min and 250 Å/min respectively (it is important to check the etch rates of the buffered solutions before starting an experiment, for they are likely to change due to crystallisation of the ammonium fluoride). The results are summarised in the chart of Figure 4.32. It is interesting to observe that there is a non-linearity at the beginning of the oxidation. It can be attributed either to a difference in reflectivity before and after polysilicon undergoes the first oxidation step or to a different "transformation" rate at the beginning of the oxidation. In both cases, this can be explained by a change in the polysilicon structure. As discussed before, polysilicon is being deposited at a temperature which lies in the "forbidden zone", within which polysilicon can either be crystalline or amorphous. If it is deposited in the amorphous state, its is most likely to change its structure in the course of any subsequent high temperature step. The change in structure can eventually mean a change in reflectivity and the formation of crystallites is likely to affect the oxidation rate.

In order to avoid implants with very shallow angles, it was decided that the final height of the walls would be in the order of 5000 Å. Apart from the oxide layer converted from polysilicon, there were already two layers which count towards this final thickness: the 300 Å pad oxide and the 900 Å silicon nitride. For the real process, the decision was to deposit a 2000 Å layer of polysilicon, which, after total oxidation, would result in an oxide layer with a thickness in the order of 3500 Å after a 2.5 hour oxidation. The exact height of this barrier is not important because the LDD lengths can be adjusted by changing the implant angle. What is important, however, is to assure that a good height uniformity is achieved across a wafer. The uniformity was verified by using a test wafer and growing the multi-layer structure on it (i.e. pad oxide, Si_3N_4 and converted oxide). The thickness of the multi-layer structure was mapped using an automated

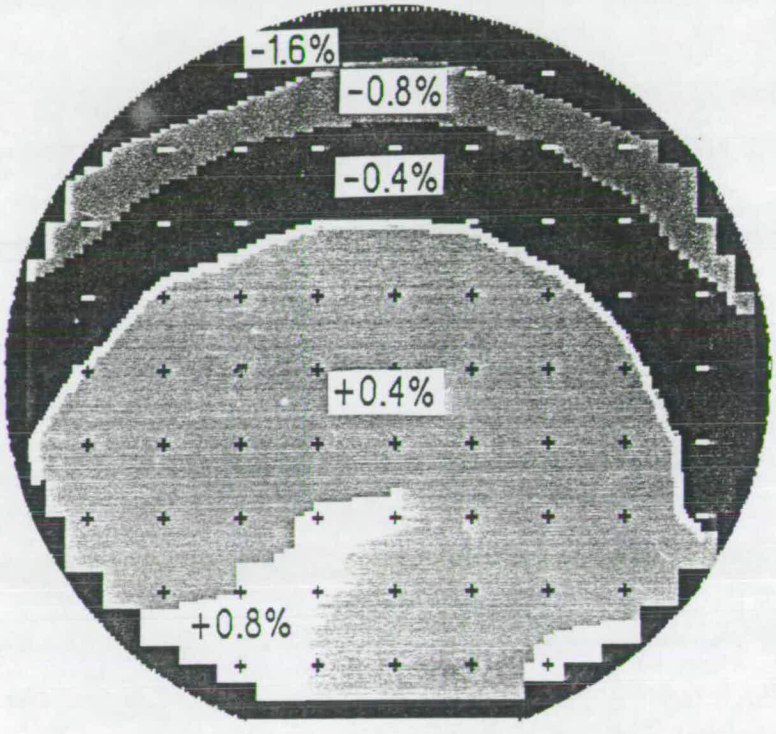


Figure 4.33: Ellipsometry results of the barrier formed by SiO_2 (converted from polysilicon)/Si_3N_4/pad SiO_2 showing the variations in uniformity.

ellipsometer, resulting in the plot of Figure 4.33.

The patterning of the implant walls was carried out in two steps: the 15 minute oxide etch in a CHF_3 : He plasma and the 3.5 minute nitride etch in CF_4 : H_2 plasma. Notice that since both plasma reactions involve fluorinated gases they can be performed in the same chamber. After those steps, the wafers had to be inspected in order to determine whether both the converted oxide and the nitride layer had been really cut through. The strategy was to measure the oxide thickness of the active regions using a spectrophotometer. If there was any residue of nitride the equipment would be unable to provide a reading within the oxide range. The obtained readings were in the order of 50 Å and below, thus indicating that not only the nitride layer had disappeared but also the pad oxide had been almost totally etched. Considering that the spectrophotometer readings for very thin oxides is not really reliable, a further confirmation was carried out by oxidising the sample and measuring it again. It showed the growth of a 300 Å oxide after a 5+2.5+5 minute (dry/wet/dry) oxidation, confirming that no nitride was left in the active areas around the wall.

The wafer batch was split in two: one for the fabrication of LDD devices and another for non-LDDs. In the case of non-LDD transistors there was no need for such a tall barrier, so the decision was to use only nitride (with no converted oxide on the top) for the source/drain implant definition. It was noticed, however, that photoresist tends to flow much more on silicon nitride than it does on SiO_2. Experiments with different hard-bake temperatures were attempted, but the final conclusion was that even at a temperature peak of 140°C, which proved to be the best possible choice, the photoresist flow was still quite significant. On the other hand, having in mind that the nitride etch process is relatively short (3.5 minutes), it was decided to attempt etching without hard-baking at all. This was successfully done and no reticulation effects were observed.

In order to accommodate transistor arrays with different LDD lengths in the same wafer, chips were grouped in different columns which were implanted using different angles. This was accomplished by isolating individual columns using photolithography before every change in angle. The angles were adjusted to produce LDD lengths from 0 to 0.3 μm in 0.05 μm steps. The angled implants were followed by a vertical implant to produce the LDD regions.

After all the ion implants, the walls had to be removed so that the gates could be defined. The process was tried initially on non-LDD transistors. The first step was to dip the wafers in 10% HF to get rid of any oxide on top of the Si_3N_4 layer. This procedure was found to damage the field oxide regions quite severely. This was caused by the weakening of thermal oxide by ion implantation. Indeed it is known [2] that ion implantation tends to increase SiO_2 etch rates up to three times. The samples were then annealed for 15 minutes at 950°C in a nitrogen atmosphere, not only to strengthen the field oxide but also to anneal the exposed active area so that it would not get etched by the hydrofluoric acid. This was followed by the silicon nitride wet etch using orthophosphoric acid at 165°C. Although this acid was not being refluxed, it had water added to it at the beginning of the operation. The wafers were left in that solution for a few hours with regular interruptions for monitoring the wall height. No change was detected. A plasma nitride etch was then attempted, again with no success even after 15 minutes with the power set to 1 kW. The effect was attributed to changes in the nitride characteristics by the ion implantation. Since this would not be the case with the LDD transistors (for the thick oxide layer on top of the Si_3N_4 would protect it against implantation) it was decided to try the barrier removal on the LDD transistors first. Their field oxide was protected by repeating the first photolithographic step (mask layer one), this time with negative photoresist so that the whole wafer would be protected except for the active areas. A few problems of infiltration at the edges of the active areas were noticed (probably caused by a bad resist adhesion to the steep profile). The wafers were etched for short intervals (usually 30 seconds each) in 4:1 buffered HF and had their profiles plotted after each cycle (Figure 4.34). The wet nitride etch was then carried out in the same conditions described earlier in the case of non-LDD transistors. The same problem observed before was found here: the nitride layer did not seem to get etched.

Before trying any new experiment, an answer had to be found for the fact that Si_3N_4 did not seem to get etched as expected. One of the difficulties was the lack of a more specific diagnostic tool to allow a fast determination of the composition of a given layer. A possible explanation for the phenomenon was that silicon was getting etched by H_3PO_4. Depending on how fast this attack was, it could eventually compensate the etching of the nitride layer, resulting in a constant step height. The only available way to test

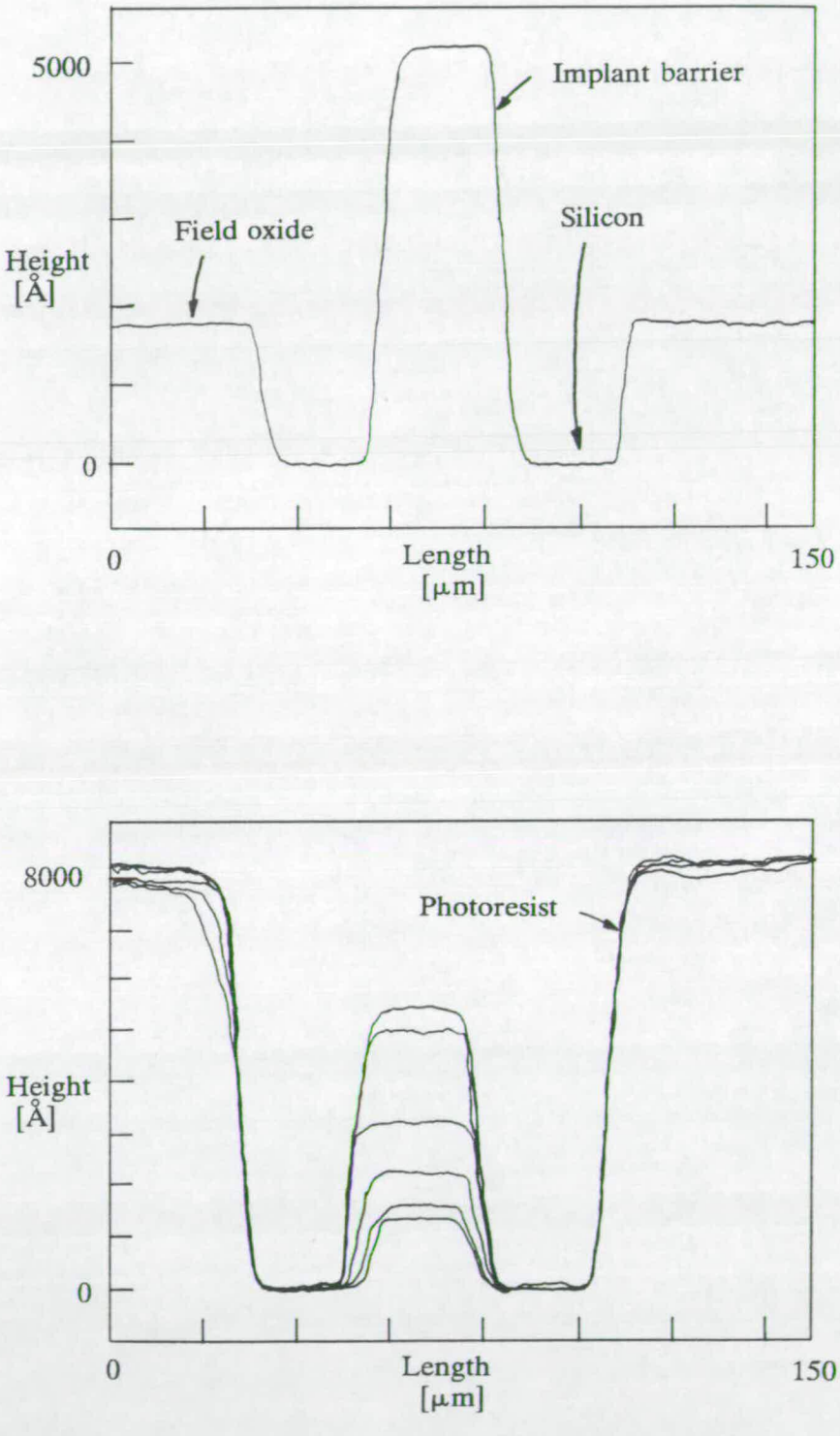


Figure 4.34: Sequence of stylus measurements obtained before protecting the field oxide with negative photoresist (above) and at different stages of the top oxide removal of the $SiO_2/Si_3N_4/SiO_2$ implanting mask barrier (below).

this hypothesis was to take one of the wafers and etch its oxides completely. After determining its profile, this wafer would then be oxidised and profiled again. This would confirm whether or not the remaining wall was constituted by nitride. This experiment resulted in a profile exactly identical to the initial one, thus indicating that the wall which was still left was in fact a step in silicon, most probably due to an attack by orthophosphoric acid. The dissolution of Si_3N_4 by H_3PO_4 is a poorly understood mechanism in which several different radicals come into play. The fact that silicon was heavily doped could have made the problem even more complicated. Nevertheless, there was still one way to avoid it: to protect the silicon surface with a thermally grown oxide. SiO_2 normally has a etch rate greater than Si, but considering the high etch rate that was being observed in doped Si the possibility of protecting it with oxide seemed plausible. A wet Si_3N_4 etch was performed and the protecting SiO_2 regions remained intact, resulting in a perfectly planar structure. When the SiO_2 layer was removed, however, there was still a step present. This step was much smaller than what had been obtained before, but still big enough to make any further electrical tests unreliable.

The residual wall, or step, in the active area could have been caused by several different effects, or even their combinations. The big difference in doping between the wall and the region around it can cause a preferential etch, as it was indeed confirmed. The plasma reactions could also have etched the bare silicon. The defects caused by ion implantation could also have caused a faster consumption of silicon during oxidation in as much as polysilicon is consumed at different rates according to its granular structure.

4.4. Conclusion

The discrepancies found in the gate offset sequences of the first batch were found to be generated mainly by rotation errors during photolithography. This could be observed in more detail in the course of the second fabrication. After this more accurate survey, it was found that although sometimes there were big differences between vernier results in the first fabrication, the errors within local groups of transistors in a small region of a chip would be acceptable.

The LDD length of the offset-gate LDD devices produced in the first process could not be uniquely measured, because of the influence of the offset gate. This will be

addressed in detail in Chapter 5. The difficulties found in the removal of the masking wall for the LDD angled implant in the second fabrication process led to the conclusion that the device surface would be severely compromised. This fabrication was therefore abandoned.

References

1. Grovernor, C. R. M., *Microelectronic Materials,* Adam Hilger, Bristol, 1989.
2. Ghandhi, S. K., *VLSI Fabrication Principles,* Wiley, New York, 1983.
3. Trapp, O. D., Blanchard, R. A., Lopp, L. J., and Kamins, T. I., *Semiconductor Technology Handbook,* Technology Associates, Portola Valley, 1985.
4. Stevenson, J. M. T. and Gundlach, A. M., "The application of photolithography to the fabrication of microcircuits," *J. Phys. E: Sci. Instrum.*, vol. 19, pp. 654-667, September 1986.
5. Sze, S. M., *VLSI Technology,* McGraw-Hill, Singapore, 1988.
6. van Gelder, W. and Hauser, V. E., "The etching of silicon nitride in phosphoric acid with silicon dioxide as a mask," *J. Electrochem. Soc.*, vol. 114, pp. 869-872, August 1967.
7. Cass, T. R. and Reddi, V. G. K., "Anomalous residual damage in Si after annealing of "through-oxide" arsenic implantations," *Appl. Phys. Lett.*, vol. 23, pp. 268-270, September 1973.
8. *Eaton 8010 Wafer Stepper User's Manual*, Mountain View, May 1983.

Chapter 5 - Device Testing

5.1. Introduction

In any fabrication process the geometric characteristics of the resulting devices deviate from the original design. However, the deviations are usually relatively well behaved and independent of each other, depending mostly on process parameters. Once determined, they are almost constant over a wafer or even a whole batch. The channel length, for instance, can deviate from its designed value in a self-aligned device because of photolithographic errors and source/drain side diffusion. The difference between the designed and the real values can be easily obtained and can be considered the same at least for different devices on the same wafer. In the present work at least two new variables were introduced: the gate offset and the LDD length. The results of the present work rely on the precise determination of those variables. However, both of them can vary at the same time from chip to chip and within a chip. Therefore special attention had to be devoted to the unique determination of those variables before degradation tests could be carried out.

Special care also had to be taken in the implementation of the degradation assessment techniques. Some of the test parameters (such as the charge pumping current) scale down with the device dimensions, making them more difficult to measure. Additionally, considering the number of different techniques available, the claims for and against each which have been published in the past and the variety of testing conditions under which they were carried out, it was quite natural to evaluate each separately before use.

5.2. Initial assessment

Before any other tests, a thorough verification of the integrity of the devices was carried out. The main aspects which were investigated were: a) junction integrity; b) oxide integrity; c) punch-through. This was done in a very simple way, using a manual

prober and a Tektronix 577 curve tracer. It has been noticed in the past that curve tracers are far more suitable for breakdown analysis than the usual automated equipment because they allow an instantaneous observation of the breakdown process, which is absolutely necessary for the interpretation of the phenomena involved (for example, walk-out phenomena can hardly be observed without a curve tracer).

It was noticed that the junctions of the self-aligned transistors were exhibiting a destructive breakdown, i.e., once the onset of the breakdown took place, the current would increase drastically and a shorted junction would result, although the current compliance of the voltage supply had been clamped at 10 μA. This effect was not observed in the offset-gate transistors. The only plausible explanation for this effect is that aluminium spiking took place, probably because the phosphorus contact implantation was not deep enough (see the considerations for the phosphorus implant design in Chapter 4).

5.3. Description of the measurement system

The large number of devices to be tested and the variety of measurement techniques employed (some of them requiring further computation, such as the calculation of derivatives) called for the use of an automated test system. It consisted of a Hewlett-Packard HP4062B parametric test system which is constituted by four SMUs (Stimulus and Measurement Unit), a 48-pin switching matrix, a switching matrix controller and a 1 MHz capacitance meter. The SMUs can be used either as voltage sources (in which case they can simultaneously measure current) or as current sources (in which case they measure voltage). SMU 1 has a higher sensitivity and can be used to monitor currents with a resolution up to 50 fA. In all the programs which were written for performing the various measurements described in this Chapter, the SMUs were configured so that low currents were always measured by SMU 1. In addition to the parametric test system, a Hewlett-Packard HP8115A dual channel pulse generator was used for the charge pumping tests. The system was connected to a Teledyne TAC PR-51 automatic wafer prober which was placed in a dark Faraday cage. Each individual equipment was connected through an HPIB interface to a Hewlett-Packard HP9000 series 300 computer.

5.4. LDD verification

Once the transistors were tested for catastrophic failures, a verification of the presence of LDD regions took place, in order to make sure that the presence of the lightly doped regions could be detected and measured with good repeatability. This was necessary due to the fact that the LDD regions were defined using photolithography, therefore subject to eventual misalignments. The design itself tried to compensate for misalignments, so that in the event of an excessive misalignment an LDD transistor would still be produced (Figure 5.1.a). However, the length of the photolithographically produced LDD is not only dependent on misalignment but also on the under/over-exposure. If excessively under-exposed, there can be no LDD effect present at all (Figure 5.1.c).

The verification of the LDD regions consisted essentially of measuring the effective channel lengths and comparing to those of non-LDD transistors. The channel conductance method [1,2] was initially used. There are several different variants of this method, which consists of measuring the channel conductance (or resistance) of several transistors (a minimum of two) and plotting the values versus the drawn channel lengths. When this is done for different gate voltages, the lines obtained for each gate voltage all intercept in a point which corresponds to ΔL. In our case we start by plotting the points corresponding to two different channel lengths (1.0 and 1.5 μm) of non-LDD transistors using a fixed voltage. The line obtained from those two points is then used to find the "altered channel lengths" (due to the presence of the LDD regions) of the LDD transistors from their measured conductances (or resistances).

In general, the drain voltage has to be kept low in order to avoid short-channel effects.[2] For the choice of gate bias two different lines of action are advocated. One is to choose a low voltage, slightly higher than the threshold voltage to make sure that strong inversion is achieved.[3-5] In this case the conductance values of different devices should be measured with a constant $V_{GS}-V_T$ voltage, which requires an accurate determination of all the individual threshold voltages. The other possibility is to choose a higher gate voltage to avoid this sensitivity to V_T,[6] which can have the adverse effect of modulating the LDD regions and cause false readings. In the present work, both low and high gate voltages were used. In the low voltage range measurements were repeated for $V_{GS}-V_T=1$, 2 and 3 V and in the high voltage range $V_{GS}=6$, 8 and 10 V. The results

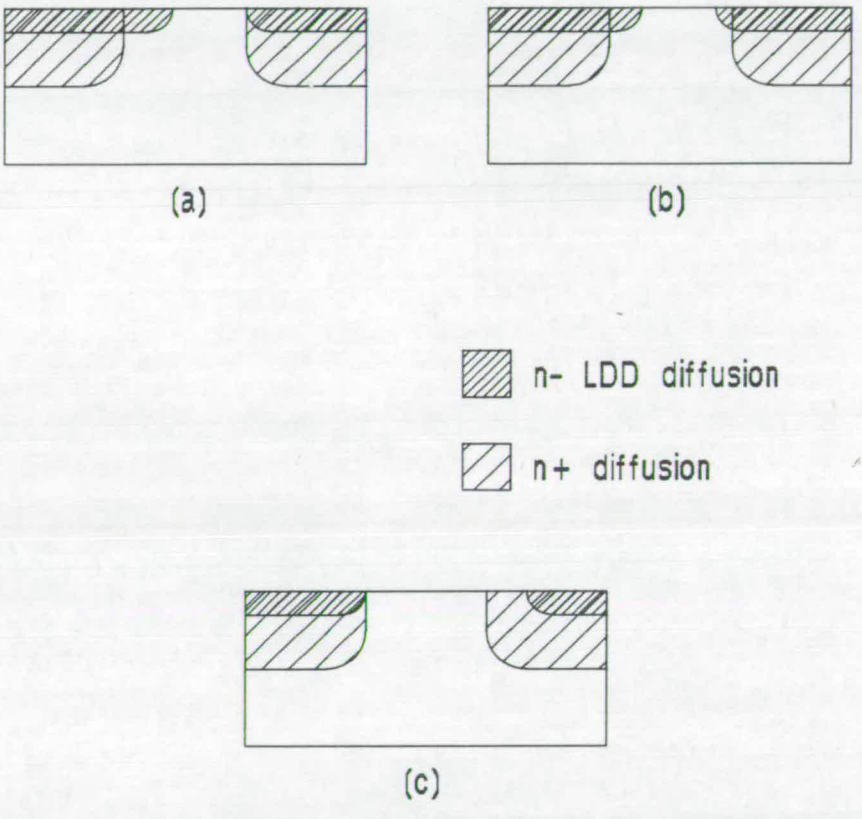


Figure 5.1: Definition of the LDD regions. b) LDD diffusions were included on both source and drain sides of the transistors. a) In the event of excessive misalignment, the overlapping LDD region on the left hand side can eventually underlap and consequently be inactive; if this happens, the right hand side will overlap. c) If the LDD photolithographic level is under-exposed, there is a risk that no LDD will be present in either side.

for groups of transistors belonging to different arrays (therefore with different drawn LDD lengths) within a chip are depicted in Figure 5.2. Notice that in all cases the variation within a group of transistors having the same drawn LDD length (points belonging to the same line) the amplitude in readings is comparable to the differences amongst lines. Although only transistors with no gaps were chosen for these measurements, the variations in gate positions were sufficient to cause the variation in resistance results. A more robust method had therefore to be found which was insensitive to those variations.

The paired V_{GS} method [7] is an adapted version of the above method and allows the direct determination of both the channel length and series resistance of LDD devices. It tries to avoid the dependence of the series resistance upon the gate bias (V_{GS} tends to modulate the LDD regions, leading to an ambiguity of series resistance and LDD length values). In our case, we are also trying to avoid this dependence, but mainly because of the variable gate position. In this method different transistors (at least two) have their channel conductances determined using pairs of gate voltages, whose difference is usually not more than 0.5 V. Provided that this difference is small, the extracted parameters are expected to vary less than 0.1 μm. The lines resulting from plotting the channel resistance versus the drawn channel length for a pair of voltages should intercept in a point whose coordinates are the series resistance and LDD length. Figure 5.3 shows the results obtained for different groups of transistors. Once again the results were inconclusive.

It is known that the gated diode current has its lower and upper edges corresponding to the flatband and threshold voltages, respectively, and its amplitude is proportional to the density of surface states.[8] Assuming that this density is uniform in fresh transistors situated within a small region of a wafer, it was thought to be possible to determine the gate length by measuring the gated diode current amplitudes. It was noticed, however, that although the variation of results amongst transistors of a given array is not too big, it cannot account for very small differences between two different arrays (as would happen with arrays differing between themselves in just 0.1 μm of LDD length), as can be seen in Figure 5.4. It is therefore inadequate for LDD length determination.

Another reliable method of channel length measurement is the channel capacitance method.[9] Firstly the open circuit capacitance is determined in order to eliminate all possible stray contributions. The gate capacitance is then measured under accumulation,

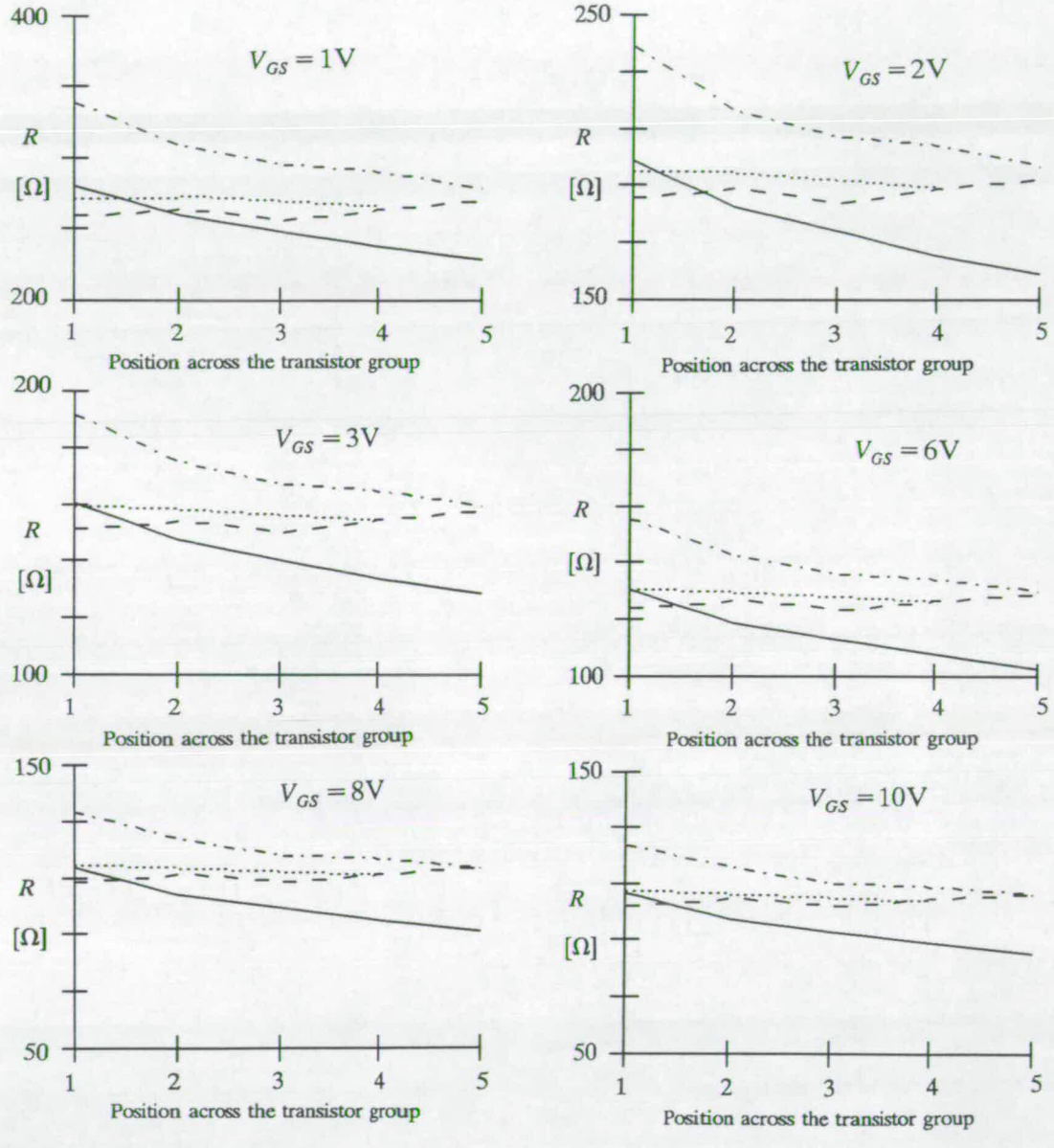


Figure 5.2: Channel conductance method. Each individual curve in a graph represents a group of transistors with the same drawn LDD length. All transistors have the same channel length (without considering the LDD regions) and have offset gates but without gaps. $V_{DS}=0.1V$.

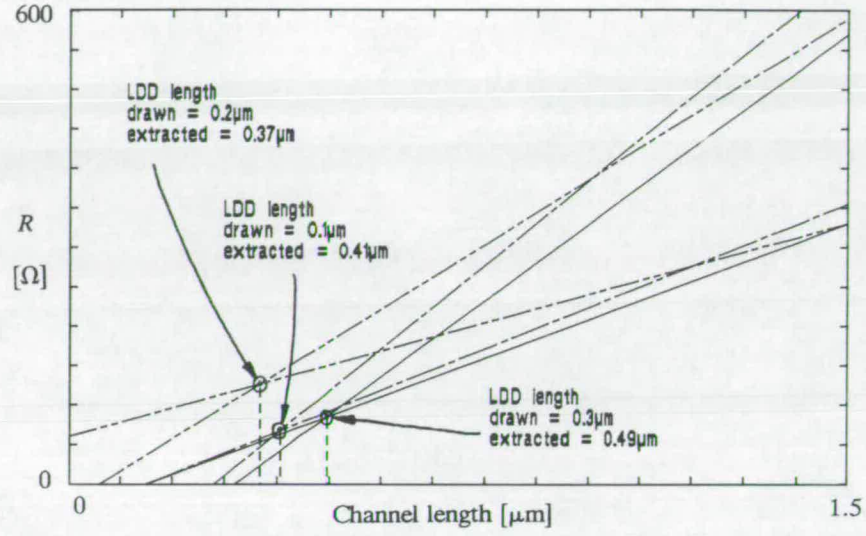


Figure 5.3: Paired V_{GS} method. Each line is drawn from two points corresponding to measurements in a 1.0 and a 1.5 μm transistor. Measurements are carried out for two different values of V_{GS}, leading to two different lines. The figure represents pairs of curves for groups of transistors with different drawn LDD lengths (each pair is depicted with a different line type). $V_{DS} = 0.1$V and $V_{GS} = 1.0$ and 1.5V.

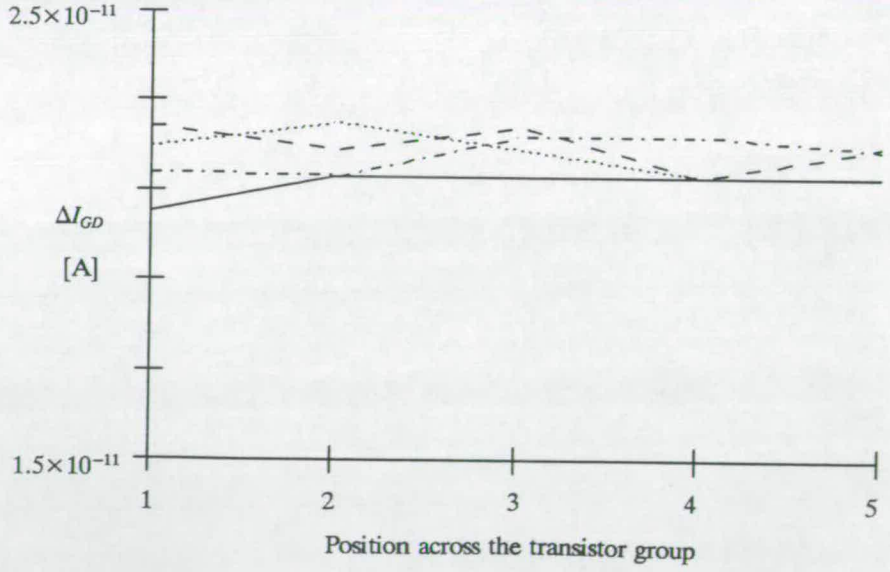


Figure 5.4: Gated diode method. Each curve represents a group of transistors with the same drawn LDD length. All transistors have the same channel length (without considering the LDD regions) and have offset gates but without gaps.

in which case the capacitance is the result of contributions of different regions of the transistor, except the active area (Figure 5.5.a). The capacitance is measured again, this time with a bias high enough to drive the active region into strong inversion (Figure 5.5.b). If we now subtract the accumulation capacitance from the inversion capacitance (and compensate for the external stray capacitances measured previously), the effective contribution of the active region results. It is important to emphasise, however, that care must be taken in order to avoid accumulating the active region excessively. If the applied accumulation voltage is too high, there is a risk that the LDD regions will be depleted and therefore not be taken into account in the computation of the active region area (Figure 5.5.c). For this reason two different voltages were used for the accumulation level: -5 and 0.5 V Figures 5.6.a and 5.6.b show the results of rows of five transistors each belonging to different arrays, each array having been designed with a different LDD length. Notice the amplitude of results within a given array, which makes the method unsuitable for our purpose of evaluating LDD lengths.

The punchthrough method suggested by Elliot for the determination of the effective channel area of a device [10] was also tried. It consists of measuring the channel capacitance whilst varying the source/drain bias. As the reverse bias across the junctions is increased, the depletion regions expand and the capacitance decreases. When punchthrough occurs, the capacitance suffers a steeper decrease. From the voltage at which this decrease occurs the width of the depletion regions can be established. However, in view of the small geometries involved, it was not possible to detect these capacitance changes with enough precision.

Finally, in an attempt to reduce the influence of the gate bias (and consequently of the gate position) even further, the parasitic bipolar transistor made up of source (emitter), substrate (base) and drain (collector) was used. It is known that the common emitter bipolar transistor gain is given by [11]

$$\beta_{d.c.} = \frac{W_e D_n N_D}{W_b D_p N_A} \tag{5.1}$$

where W_e and W_b are the widths of the emitter and base regions, D_n and D_p are the electron and hole diffusion coefficients and N_D and N_A are the emitter and base doping levels, respectively. Considering that amongst different groups of transistors with

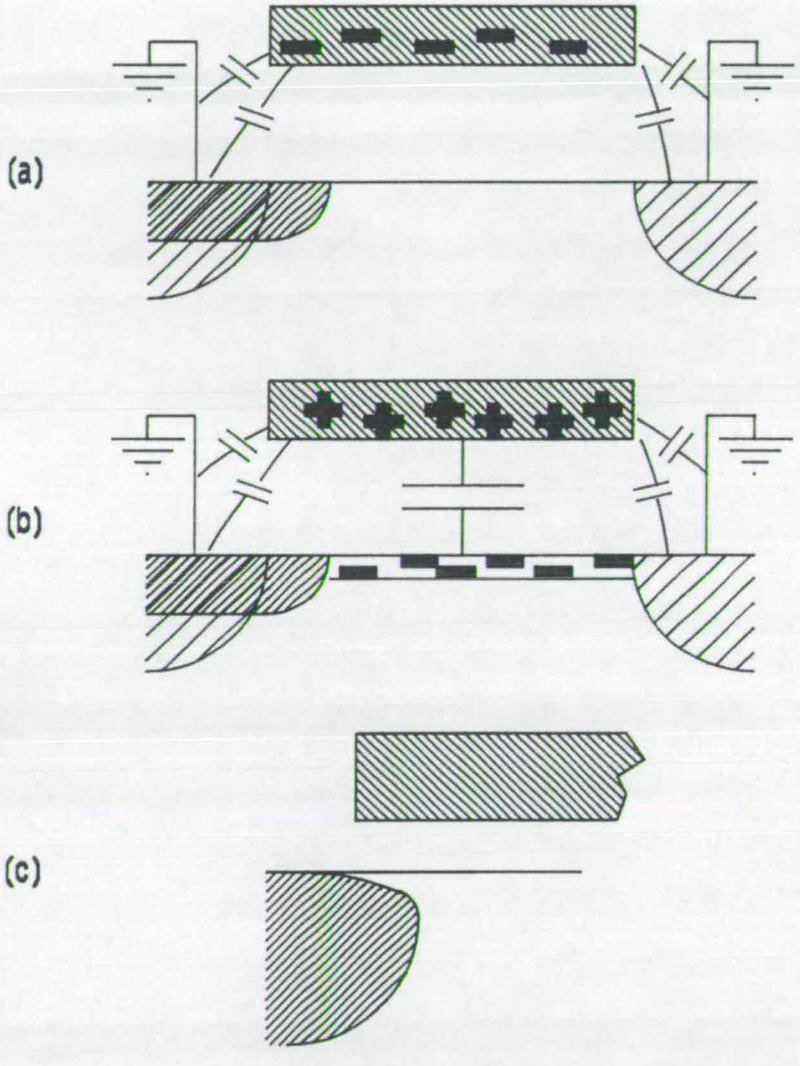


Figure 5.5: Intrinsic capacitance method. a) A capacitance value which includes all contributions except the channel capacitance is initially obtained by measuring the gate capacitance with the channel in accumulation. b) The measurement is repeated with the channel in strong inversion and the channel capacitance is computed as the difference between those two. c) If the device is too strongly accumulated for the first measurement, there is a risk that the LDD regions get inverted, therefore leading to false channel length results.

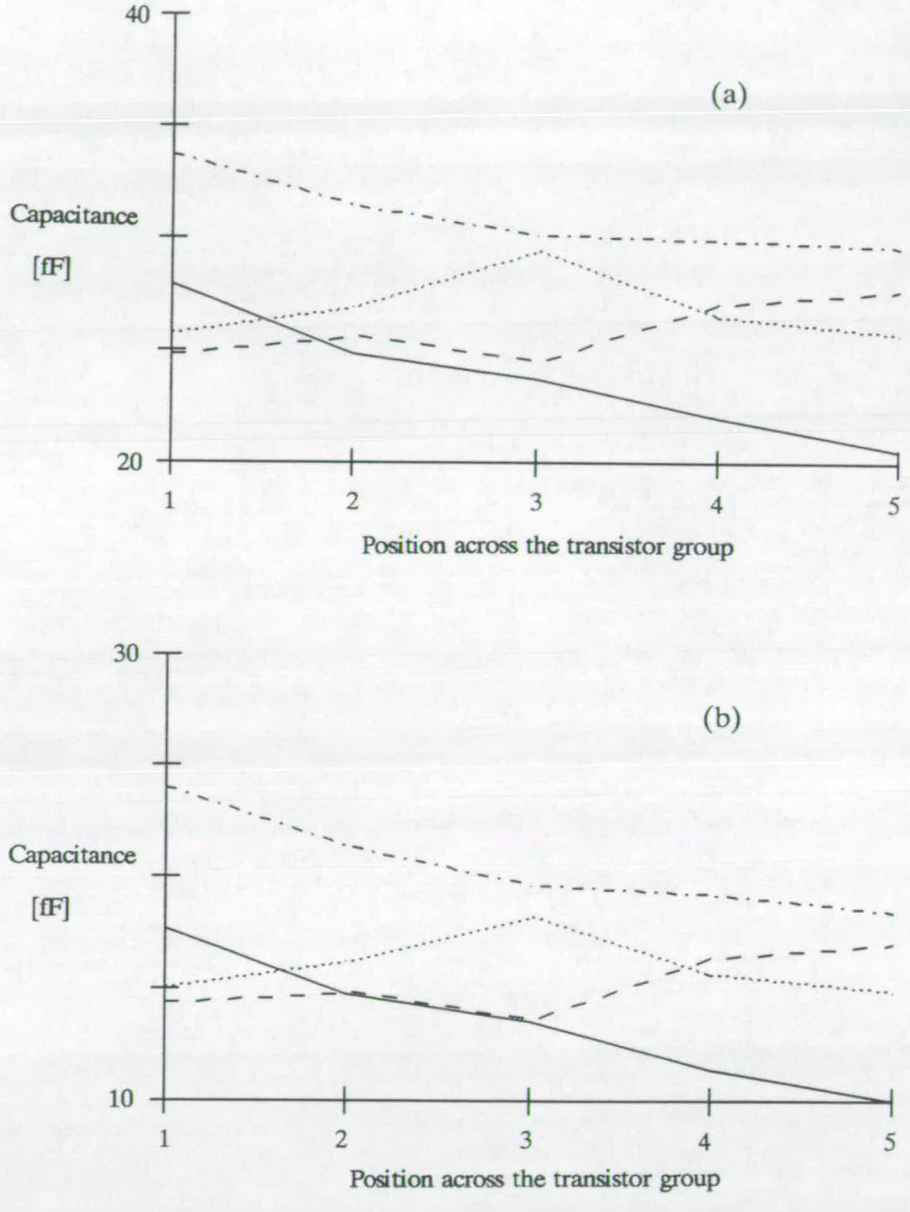


Figure 5.6: Capacitance results. Each individual curve in a graph represents a group of transistors with the same drawn LDD length. All transistors have the same channel length (without considering the LDD regions) and have offset gates but without gaps. The initial measurement was carried out with the channel depleted (a) and strongly accumulated (b). $V_{depl} = 0.5\text{V}$, $V_{inv} = 5\text{V}$ and $V_{acc} = -5\text{V}$.

different drawn LDD lengths the only difference is the channel length (which in this case becomes the base width), it should be possible to correlate the transistor gain with the channel length. A low-current gain should preferably be used, since the gains above the maximum point (also known as constriction point) will be affected by current crowding effects, which will be rather unpredictable in a device which was not designed to operate as a bipolar transistor. During the measurements the gate was left floating†. The variation within arrays was too big (Figure 5.7), probably caused by the variation in gate positions. For this reason this method also turned out to be unsuccessful.

None of the techniques above was totally independent of the gate position. Additionally, no good correlation was found amongst them, which either indicates that they were not sensitive enough to account for the presence of the LDD regions or that the dimensions of such regions suffered a process-related distortion (which could even mean the total absence of LDD overlap), therefore not corresponding to what was originally designed. Such distortions would also lead to random variations in LDD length within an array whose LDD lengths were drawn to be constant. Even if it was possible to use one of the techniques above to find our LDD lengths the very possibility of having variations within a group would make such a technique unusable unless it was possible to determine the LDD length in an individual device. This design suffers particularly because of its relatively large frame size (1 cm^2), which is subject to eventual edge aberration and image rotation during the photolithographic steps. Such problems are probably negligible in conventional designs, but become crucial in designs like this where one relies on the presence of regularly incremented geometrical variables. As it will be seen in the case of evaluating gate offsets, such a distortion/rotation indeed happened.

5.5. Gate offset verification

The gaps in different transistors within a given array of transistors had to be assessed by a method which: a)allows a fast (possibly automated) determination of their sizes; b)does not significantly depend on other variables; c)does not introduce

† All current/voltage sweep measurements performed in the course of this work were preceded by grounding all device terminals in order to remove any residual electrostatic charge. This was particularly important for the gated-diode current measurements, which are discussed later in this chapter.

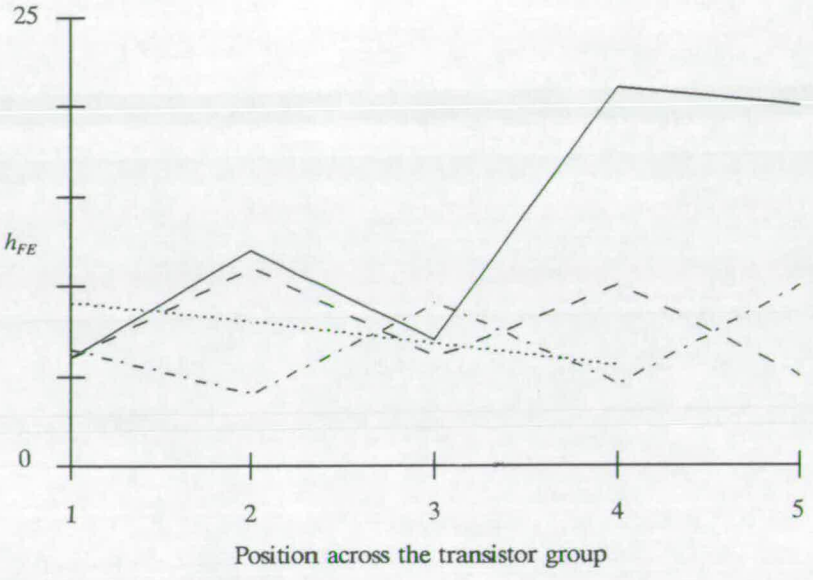


Figure 5.7: Common emitter gain method. Each curve represents a group of transistors with the same drawn LDD length. All transistors have the same channel length and are non-gapped. $V_{CE} = 1\text{V}$, $I_C = 0.5\text{nA}$.

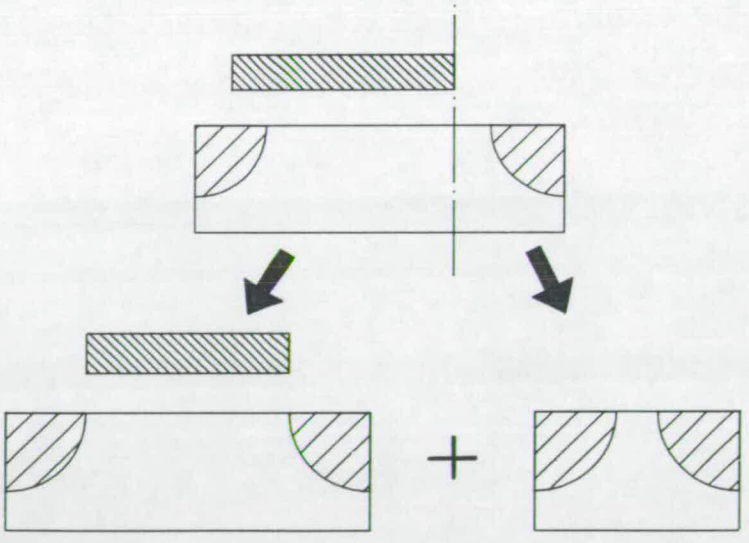
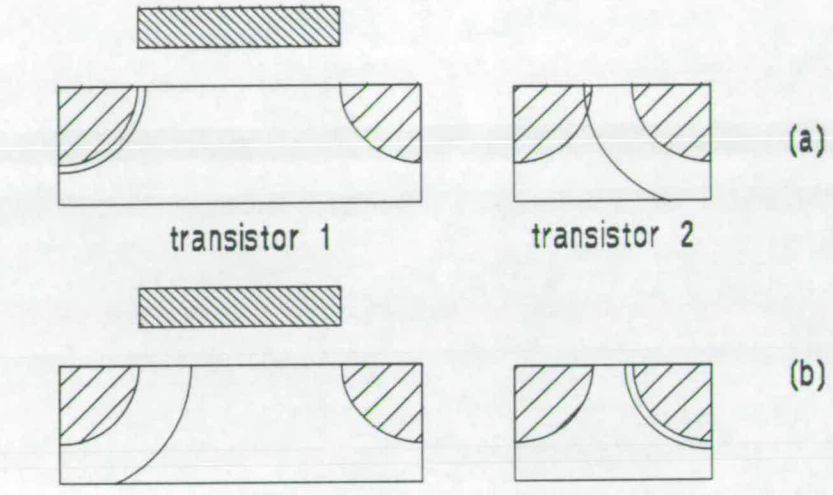


Figure 5.8: The gapped transistor seen as two transistors in series: one is a non-gapped transistor and the other is a transistor without a gate.

degradation (some of the gapped transistors can be very sensitive to certain types of degradation, as will be seen). Additionally, like any other electrical measurement method, the ideal situation would be to determine gaps unequivocally, i.e., a single transistor has to be tested to give its own gap value. Another possibility is to compare the measurements with simulated results. This has been avoided in the present work because such a comparison can only be successful when 1) the simulation models are accurate enough, 2) the doping profiles are very well known and 3) the characteristics of the device surface are well determined and input to the simulator. If any of those conditions is not satisfied, the correlation can be totally misleading.

The comparison of subthreshold characteristics to determine gap sizes has been tried successfully before. [12] For that matter, the gapped transistor can be seen as consisting of two different transistors connected in series (Figure 5.8). Transistor 1 has a totally covered channel, whilst transistor 2 has no gate on it. Let us analyse initially the situation of a small gap transistor as depicted in Figure 5.9. When the closest junction to transistor 1 is biased as source and the other junction (the closest to transistor 2) as drain, the space charge region of the source junction is totally covered by the gate of transistor 1. Although there is no gate to control the conduction of transistor 2, the depletion region extends across the entire length of its channel and therefore corresponds to the situation of a punch-through (Figure 5.9.a). If source and drain biases are now swapped, the space charge region closest to transistor 2 will have the source bias applied to it, and therefore will be much shorter than in the previous case. It will not go across the whole of the channel of transistor 2, whose conduction will depend on the lateral gate field (Figure 5.9.b). The effect of this lateral gate field is small compared to the usual vertical field, so that transistor 2 behaves as if it had a gate whose bias was lower than the voltage applied to the gate of transistor 1. If we then compare the overall situation between normal and reverse biases, we see that transistor 1 was affected to a minimum extent (the only difference between forward and reverse biases is that its effective channel length changed slightly) but transistor 2 went from punch-through to a weak gate bias. The consequence is that there is a significant difference between the subthreshold currents for the forward and reverse cases (Figure 5.9.c).



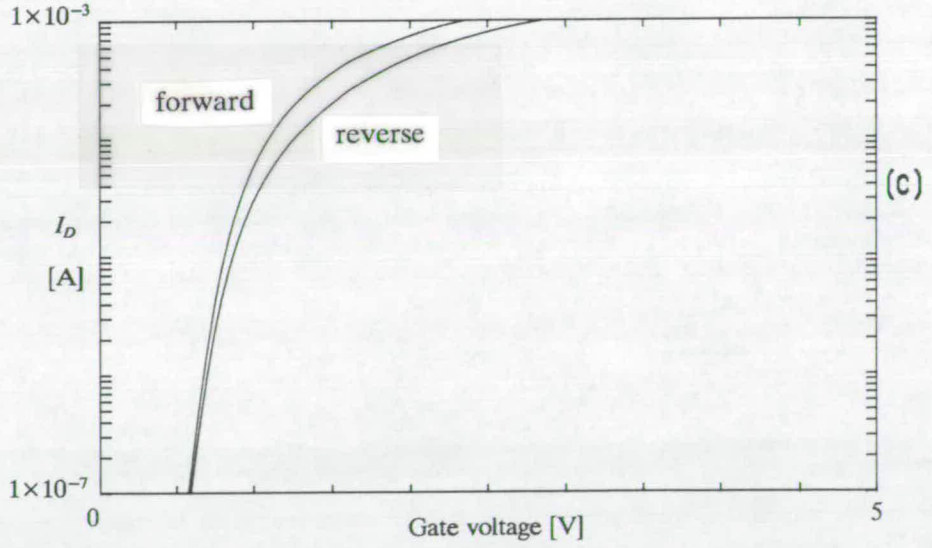


Figure 5.9: Direct and reverse subthreshold current measurements of gapped transistors. The gapped transistors are seen as two transistors connected in series as in Figure 5.8. a) Transistor 2 is under punch-through, so the channel conduction is depending on transistor 1 only. b) When the source/drain biasing is reversed, the conduction in transistor 2 will depend on the lateral gate field, resulting in poor conduction compared to transistor 1. c) Subthreshold curves resulting from both configurations.

It is interesting to notice that if the gap is large compared to the drain space charge width, both forward and reverse curves will be affected if compared to a non-gapped transistor (Figure 5.10). Therefore for a sequence of devices with increasingly big gaps, the difference between forward and reverse curves will increase with gap size until the gap size becomes comparable with the drain space charge width; at that point there will be a gap in both the forward and reverse situation and both curves will appear "bent" if compared to a non-gapped transistor. Figures 5.11 shows subthreshold results of transistors with gaps ranging from zero to 0.3 μm with 0.05 μm increments.

In comparing different subthreshold curves of different transistors within an array, it was frequently observed that there were inconsistencies. For example, Figure 5.12 shows subthreshold results of two transistors situated in two different rows. Their gap sizes were designed to be the same, with different gate lengths. The same inconsistency was observed more or less pronouncedly in many different arrays. The arrays which were tested for this purpose were the non-LDD arrays, situated at the upper-right corner of the test chip. It was therefore inferred that the cause of such a problem was a field distortion (probably some form of spherical aberration) of the DSW (direct step on wafer) machine used for the photolithographic process. Such distortions, although rare, are likely to occur in the event of an accidental change in the environmental conditions of the optical elements. It was later verified that what was in fact causing this anomaly was an error rotation alignment carried out just before starting the steppering of a wafer. Details of it can be found in the discussion of the processing of the second batch (Chapter 4). The conclusion was that groups of gapped transistors could only be used with confidence within small regions of an array. Considering that in most of the cases a suitable range of gaps could easily be found within groups of about three rows and that a tolerance of less than 0.05 μm could be achieved in such groups, this was considered not to be detrimental.

We have seen that the subthreshold current comparison technique has the disadvantage of having to depend on comparisons among groups of transistors rather than producing results straight from single transistor measurements. However, it can run automatically and is not prone to the influence of other variables (particularly because it is usually being performed within groups of transistors among which the only variable is

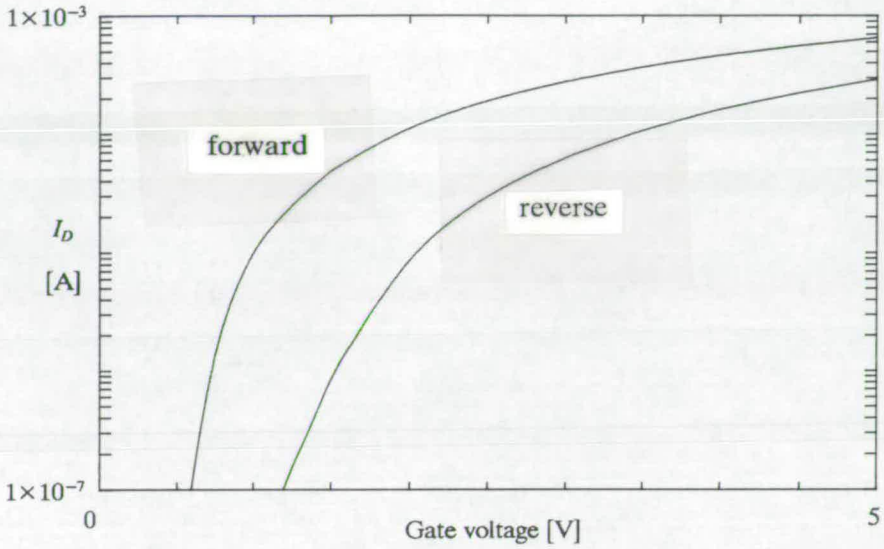
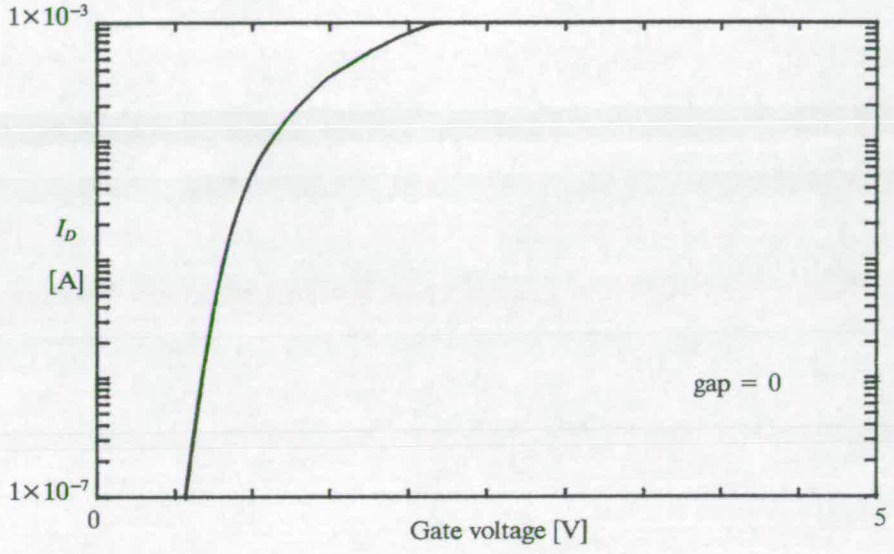


Figure 5.10: Subthreshold curves of direct and reverse configurations of a transistor with a large gap (0.25 μm).



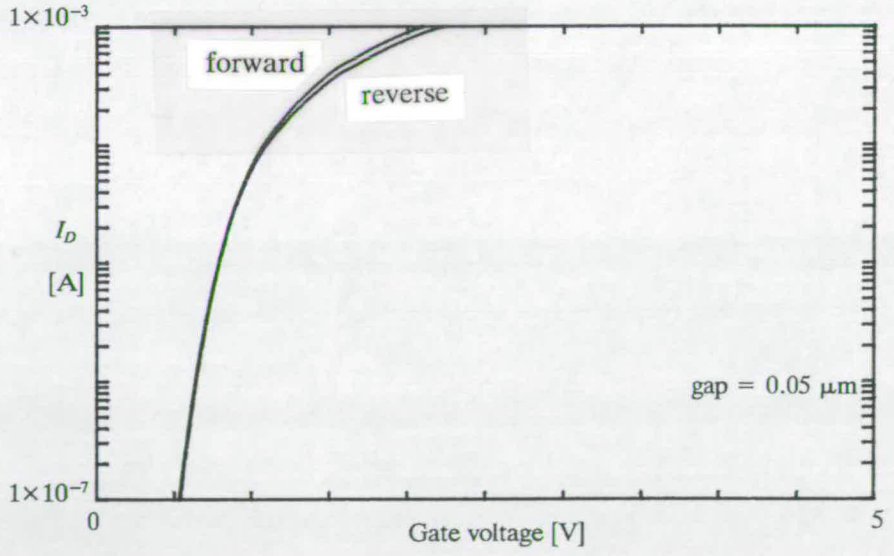
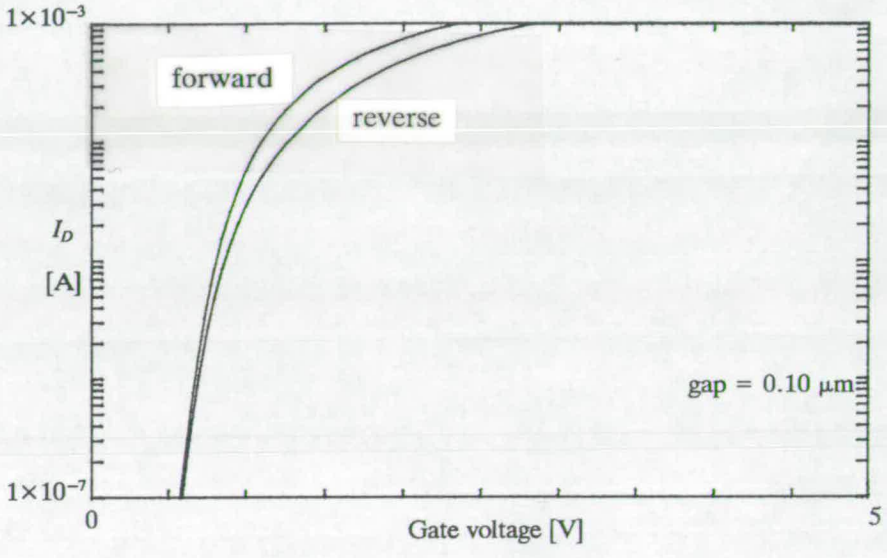


Figure 5.11: Subthreshold current comparison between the normal and reverse bias of transistors with gaps ranging from 0 to 0.3 μm steps. $V_{DS}=1$V, $V_{GS}=0$ to 5V and $V_{BS}=0$.



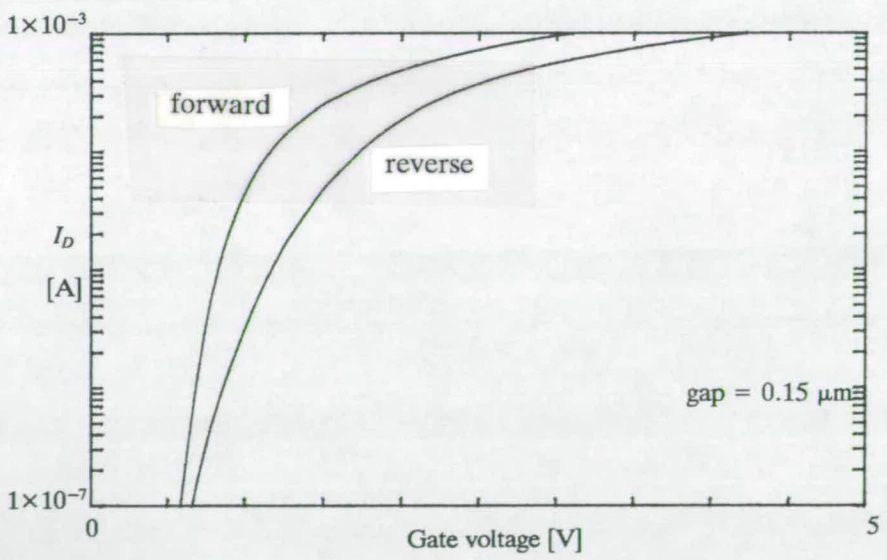
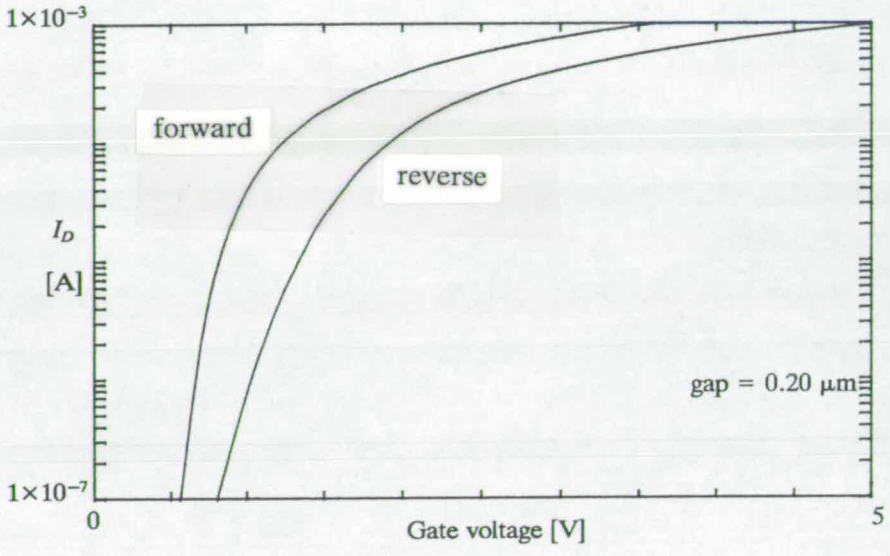


Figure 5.11 (cont'd)



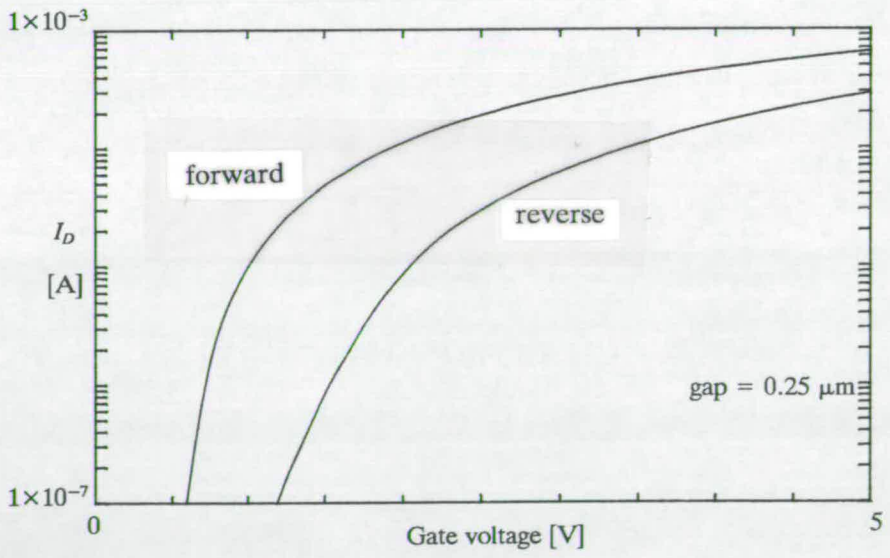


Figure 5.11 (cont'd)

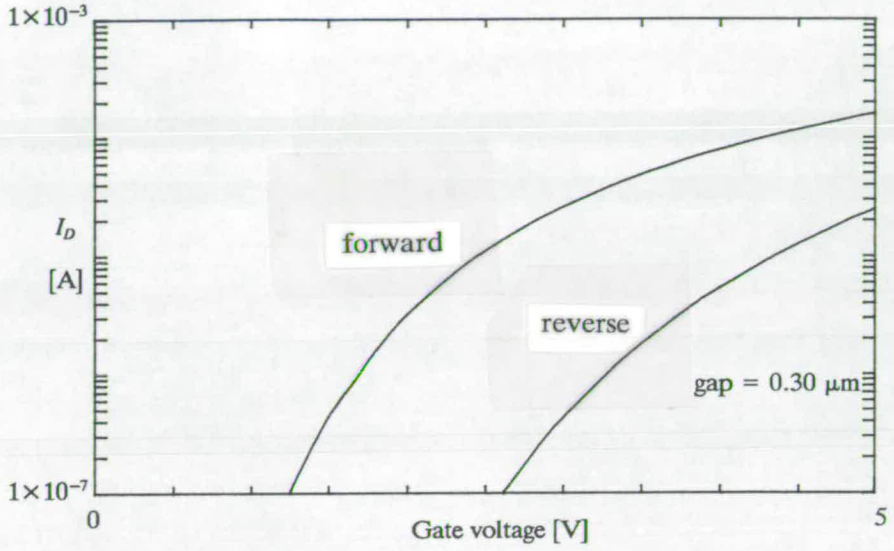


Figure 5.11 (cont'd)

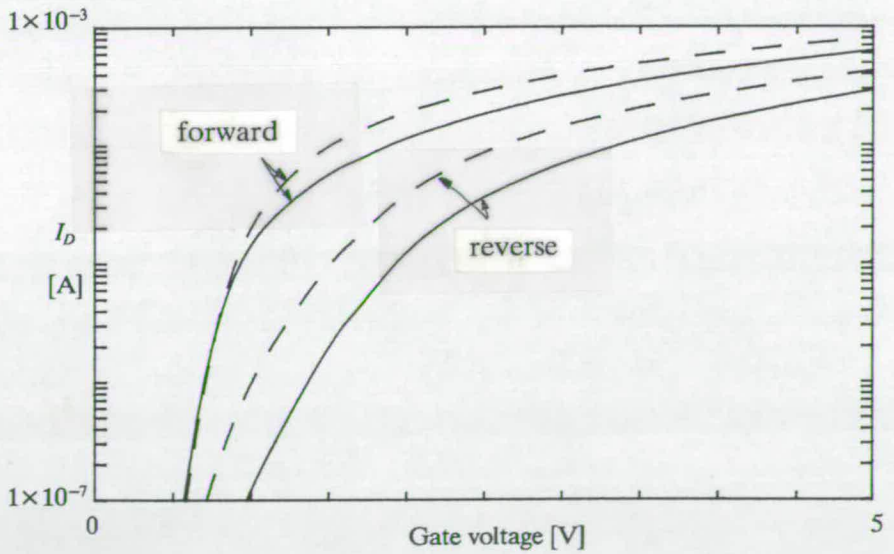


Figure 5.12: Comparison between two transistors with the same drawn gap size (0.25 μm) situated in different rows within the same chip (the forward and reverse subthreshold curves of one transistor were plotted with continuous lines and the curves of the other transistor were plotted with dashed lines).

the gate position). Another advantage is the fact that it can be executed with low integration times (due to the relatively high current level from which the onset of the current difference between forward and reverse configurations usually occurs). This factor not only increases the test speed but also avoids any eventual degradation which the test procedure itself might introduce.

Another measurement method was used as an attempt to double-check the offset results: the double-hump technique. It has been observed [13, 14] that when the gate voltage is swept from zero volts towards strong inversion the substrate current goes through a maximum (Figure 5.13). If there is a gap near the drain junction two distinct peaks (or a "double hump") are observed. As we have already discussed on Chapter 3, a non-gapped transistor with a fixed drain voltage and $V_{GS} = V_{DS}$ is driven into saturation when the gate voltage is reduced. When that happens, pinch-off will occur and there will be impact ionisation in the depletion region between the pinch-off point and the drain junction. The substrate current will be constituted by holes generated in this depletion region (as we saw in Chapter 3, some of those holes will be heated by the electric field and then injected into the SiO_2; however, this is not relevant in the present discussion). When the gate voltage is increased from this saturation condition, the pinch-off point will come nearer to the edge of the drain junction and eventually the linear regime will be reached. Impact ionisation will stop and the substrate current will be reduced to the reverse saturation current of the drain junction. In offset-gate transistors, however, pinch-off will still be present when V_{GS} equals V_{DS} (this geometry-induced saturation will be discussed in detail in Chapter 6 in order to explain the role of gaps in the degradation mechanisms). Because there is an increase in the electric field at higher gate voltages, there is a second peak in substrate current. The double-hump effect has been actually suggested [13] as a means to monitor the amount of gapping in transistors. Figure 5.14 shows substrate current (double hump) results obtained from the same transistors whose subthreshold current results are depicted in Figure 5.11. The curve remains almost unchanged for drain gaps from 0.10 to 0.20 μm. We therefore conclude that the double hump characteristic can indeed detect gate offsets but is not sensitive enough to allow its quantification. An additional problem is the need for relatively high gate voltages, which can be a problem if degradation during tests is to be avoided.

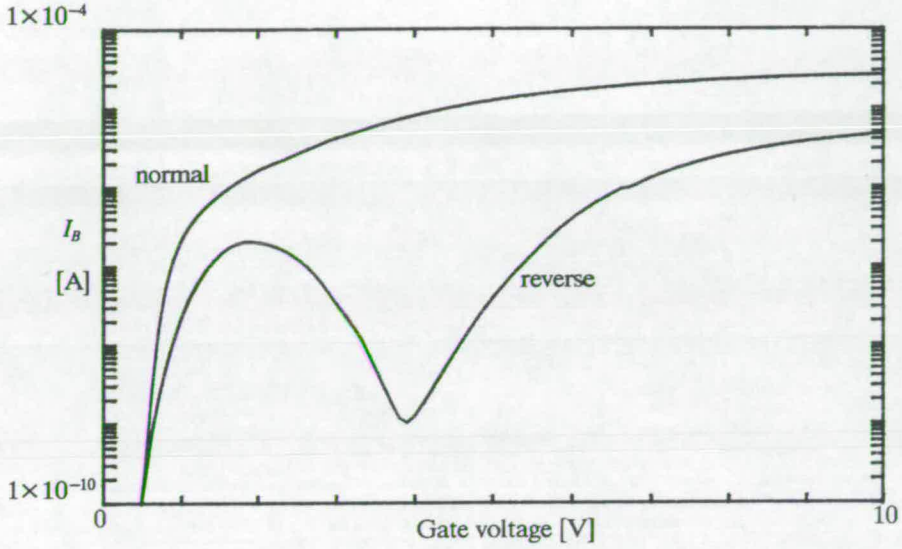


Figure 5.13: Substrate current comparison between the forward and reverse biases of a gapped transistor. $V_{DS} = 5\text{V}$, $V_{GS} = 0$ to 10 V and $V_{BS} = 0$.

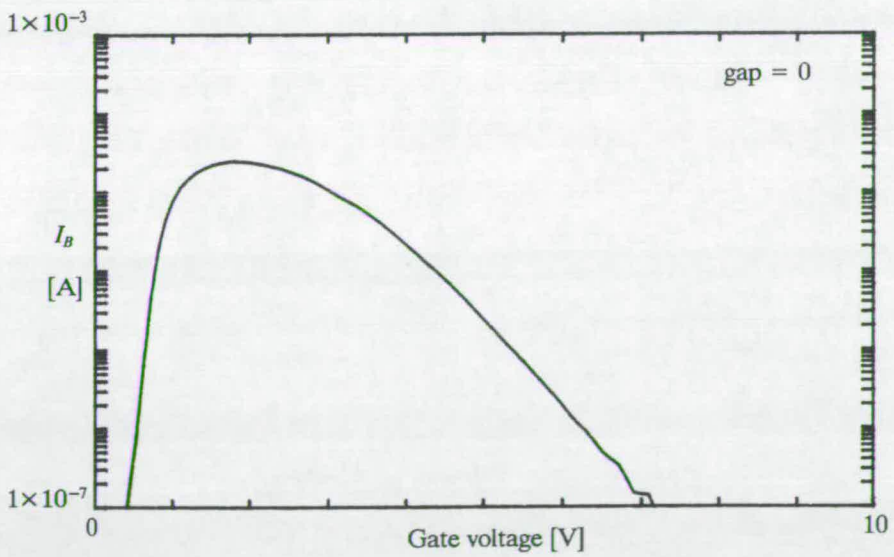
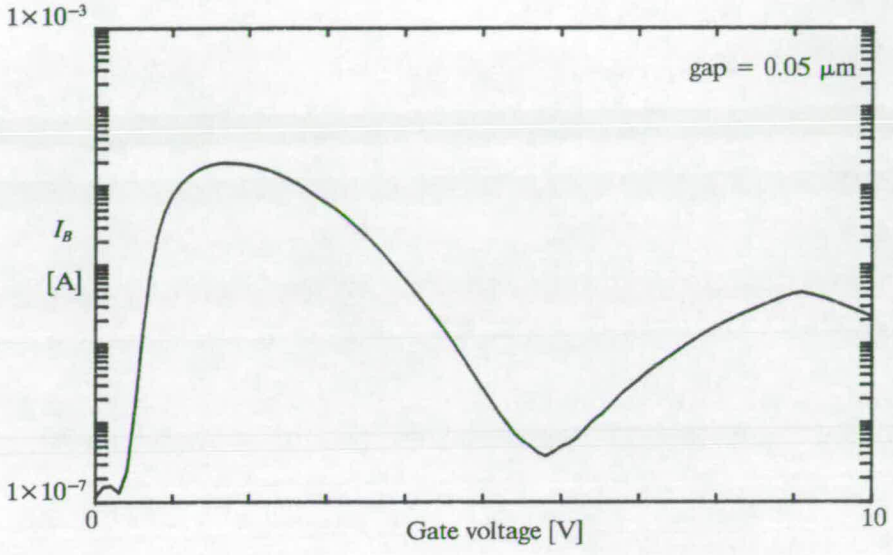


Figure 5.14: Double hump characteristic of the same gapped transistors whose subthreshold comparisons are depicted in Figure 5.11. $V_{DS} = 5\text{V}$, $V_{GS} = 0$ to 10V and $V_{BS} = 0$.



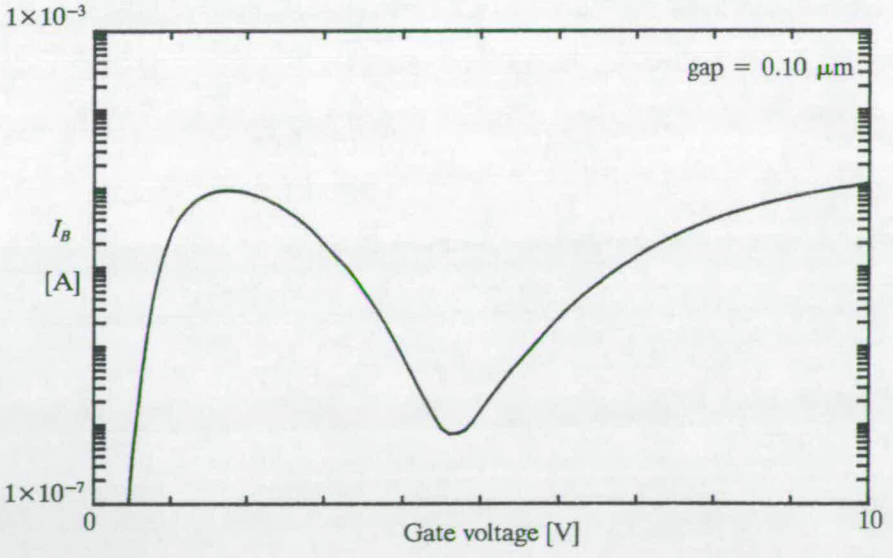
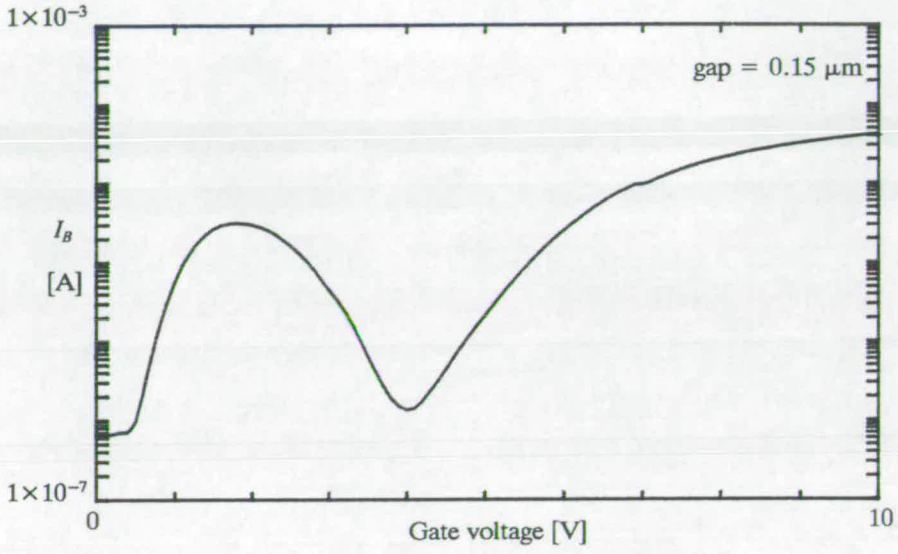


Figure 5.14 (cont'd)



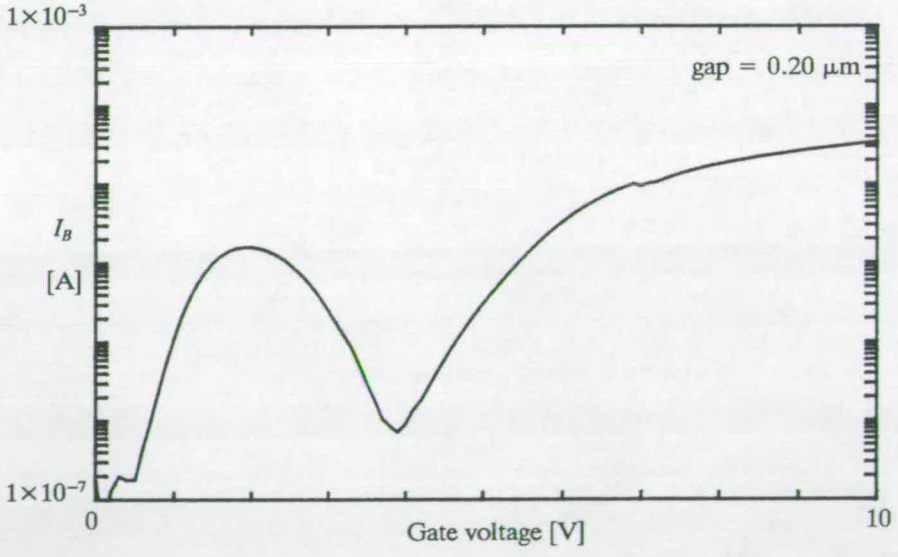
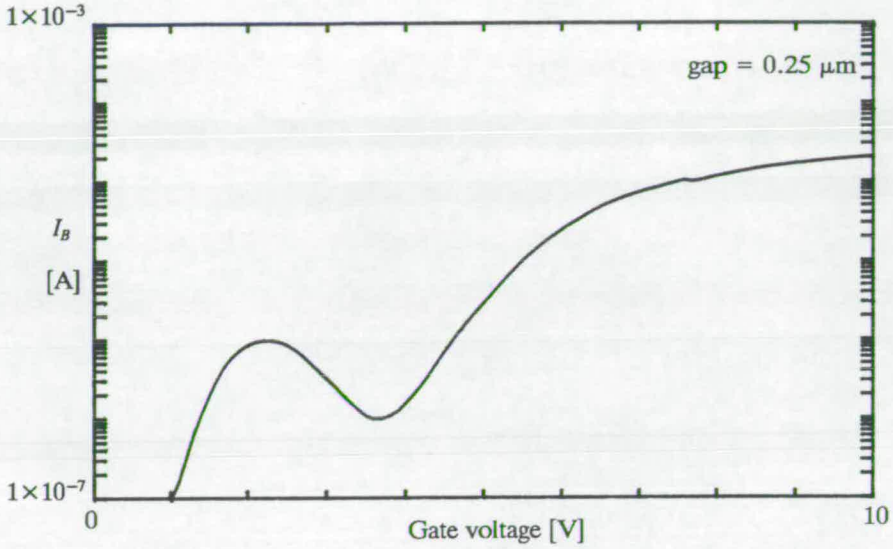


Figure 5.14 (cont'd)



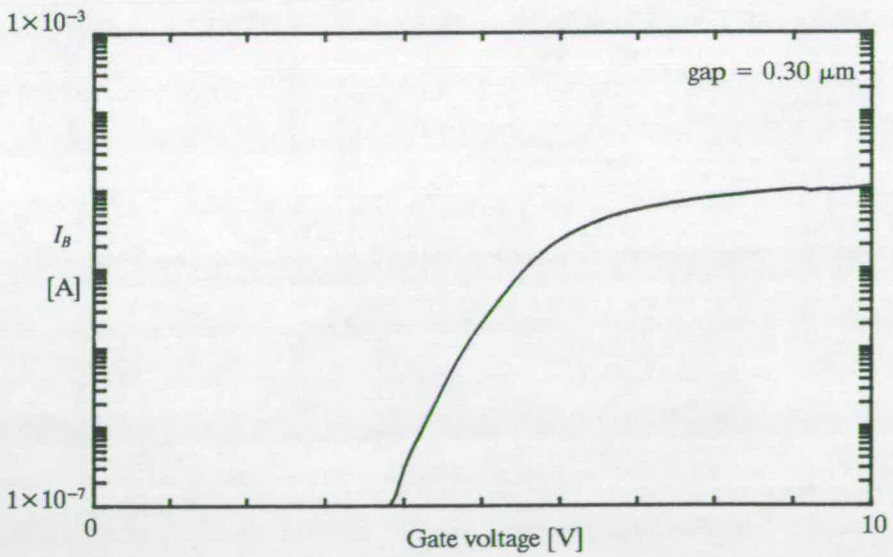


Figure 5.14 (cont'd)

5.6. Implementation of the charge pumping method

Initially it was tried to measure the charge pumping current using a voltage pulse of varying amplitude, as suggested by Brugler [15] in the original work on the charge pumping technique (in our case the low level was kept constant at -3 V and the high level was swept from 0 to 5 V). The source and drain junctions were both biased at different voltages: 0, 1 and 2 V. This basic method was used in order to make it possible to diagnose any eventual faults in the implementation of the charge pumping technique. Tests were carried out using 2 μm, self aligned transistors. It was noticed that there was a considerable current offset, which was immediately attributed to the calibration of the HP4062 SMUs. The initial curves are depicted in Figure 5.15. After introducing a calibration routine preceding all the tests, there was a noticeable improvement as can be verified in Figure 5.16.

The main charge pumping measurements used in the present work follow the method originally suggested by Elliot [10] and consist of applying to the gate a voltage pulse train with constant amplitude with a varying d.c. component so that the base level goes from accumulation to strong inversion. As previously discussed in Chapter 3, the leading and trailing edges of the charge pumping current curve correspond to the threshold voltage minus the pulse amplitude and to the flatband voltage, respectively, both with their "special charge pumping meaning". [16] The use of junction voltages different from zero is usually recommended because of the adverse effect of bulk recombination, also called the "geometrical effect" in the early works on charge pumping effect.[15] This is caused by the carriers which are not fast enough to drift back to the junctions and suffer recombination at the bulk of the channel. These excess carriers cannot be efficiently removed by the built-in voltage[10] and constitute an extra contribution towards the charge pumping curve. However, it was observed before [15] that if enough "collective area" is supplied the effect can be eliminated. This can be done by making $W >> L$. In our case, all the 1.0 and 1.5 μm transistors are 25 μm wide, so the presence of the geometrical component is greatly reduced. Any residual contributions of the geometric effect would show up at very low currents only; there is no evidence of it as attested by the logarithmic graph of Figure 5.17. The absence of the effect is further demonstrated by Figure 5.18, which shows the dependence of the charge pumping curve

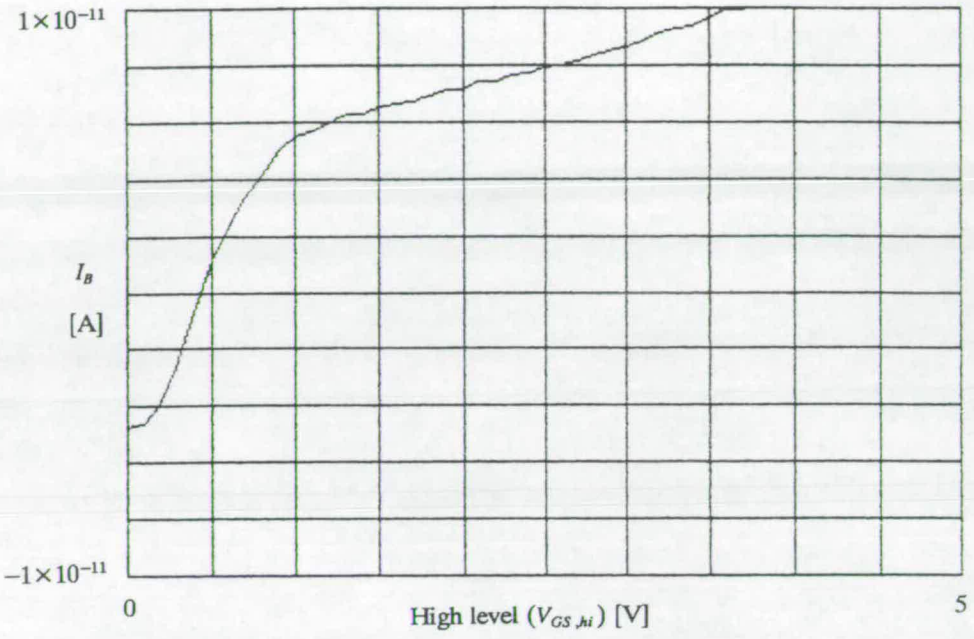


Figure 5.15: Initial charge pumping results. A 50kHz rectangular waveform oscillating between a fixed base voltage of -3V and a high level varying from 0 to 5V was applied to the gate of a 5μm transistor. Both source and drain junctions were kept at zero volts.

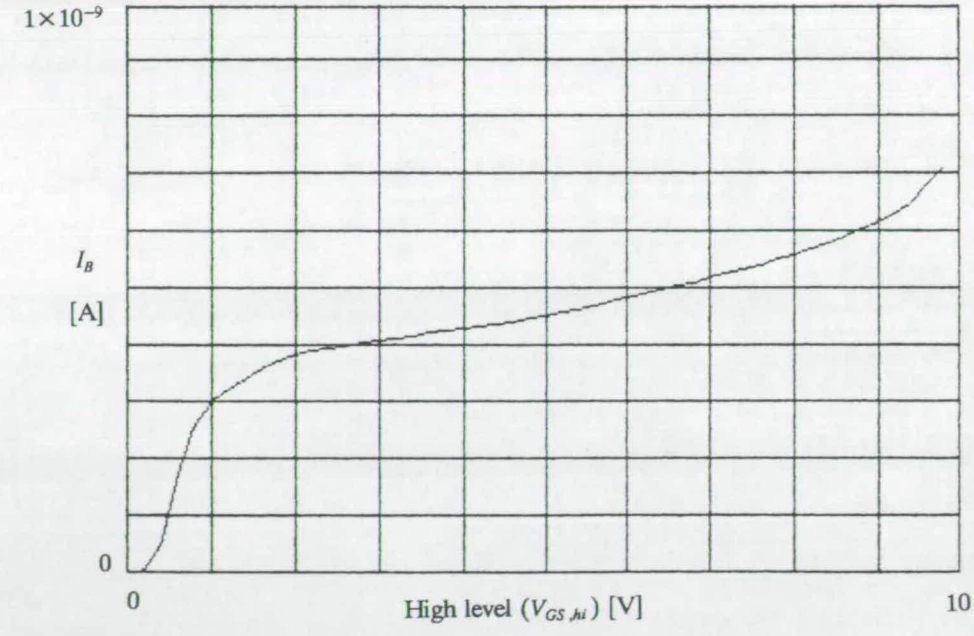


Figure 5.16: Charge pumping current measured after calibrating the test system. A 250kHz rectangular waveform oscillating between a fixed base voltage of -3V and a high level varying from 0 to 10V was applied to the gate of a 5μm transistor. Both source and drain junctions were kept at zero volts.

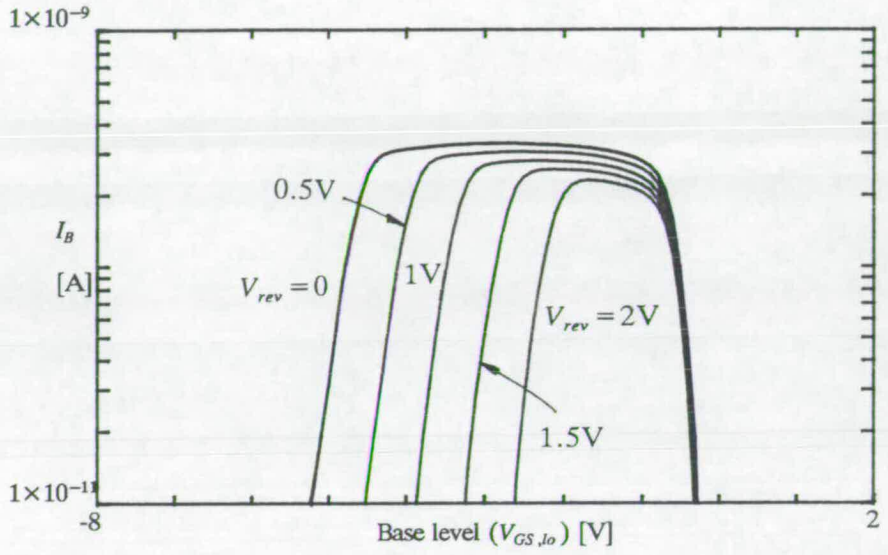


Figure 5.17: Charge pumping results for different junction reverse biases. A logarithmic y-axis was used in order to detect any changes at low currents. $f=$ 250 kHz and $\Delta V_{GS}=5$V.

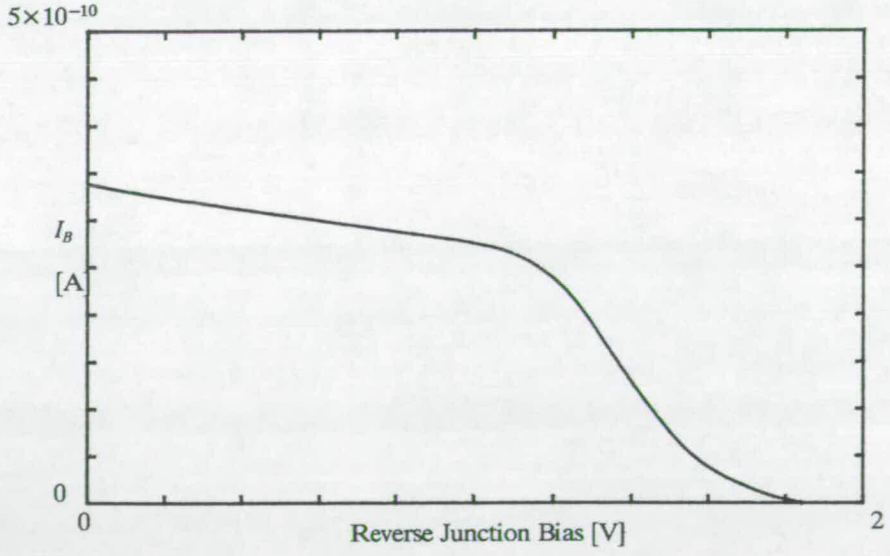


Figure 5.18: Charge pumping current versus junction reverse bias. Notice that the top region remains linear even when V_{rev} approaches zero. $V_{rev}=0$ to 2V, $V_{GS,lo}=-3$V and $V_{GS,hi}=2$V.

upon the reverse voltage (notice that the slope does not change significantly as the reverse voltage approaches zero). By setting V_{rev} to zero, one has the advantage of bringing the leading edge of the charge pumping curve to a lower voltage, thus making it easier to observe any changes in its shape.

As a further verification of the reliability of the charge pumping test routine, different transistors were tested repeatedly to see if all the curves would coincide. A typical result is shown in Figure 5.19.

5.7. Implementation of the gated diode method

In comparison with the charge pumping measurement technique, the gated diode method was much more difficult to implement, particularly in terms of assuring good repeatability. Ever since the first measurements were taken, it was noticed that the leading edge of the curve (corresponding to the flatband voltage) changed a lot if a series of repeated measurements were performed. This is exemplified in Figure 5.20.

The problem was first thought to be caused by the presence of slow states, because it also had the characteristic of being recoverable (a further measurement after a day would show a small recovery). A close look at the fabrication process for the wafers used in the initial tests revealed that they belonged to the group whose cleaning process just before gate oxidation was RCA only. Therefore it was decided that a wafer belonging to the other group (RCA clean followed by Karo's clean) also had to be tested. The results are shown in Figure 5.21. Although they were much better, it was decided to investigate this same sample a little further. The wafer was tested again after a rest period of 15 hours and a dramatic increase in flatband voltage was observed. Another five consecutive tests were carried out immediately after this first test, which showed a decrease followed by an increase in flatband voltage (Figure 5.22). The hypothesis of slow states was then discarded and the possibility of mobile ionic charges was considered. However, there were two main questions to be answered at that stage: 1) is the effect due to some device phenomenon or is it being introduced by an improper implementation of the testing routine? and 2) is the effect characteristic of this process (and more particularly of this batch)? A good step towards answering both questions was to use standard test strips of other batches of n-channel devices. The reason for using standard EMF test strips in

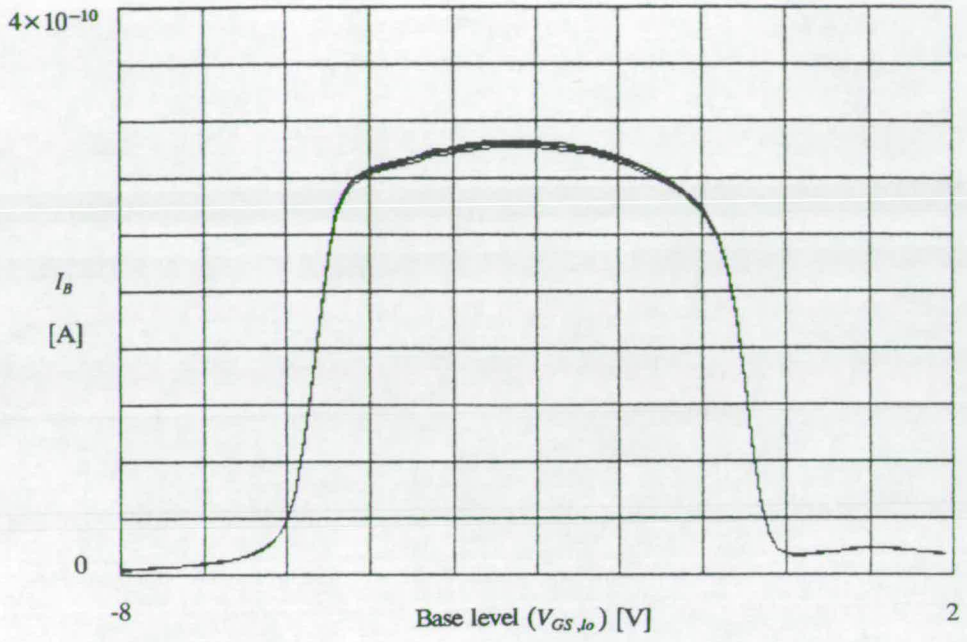


Figure 5.19: Charge pumping measurements repeated five times consecutively. $f = 250\text{kHz}$ and $\Delta V_{GS} = 5\text{V}$.

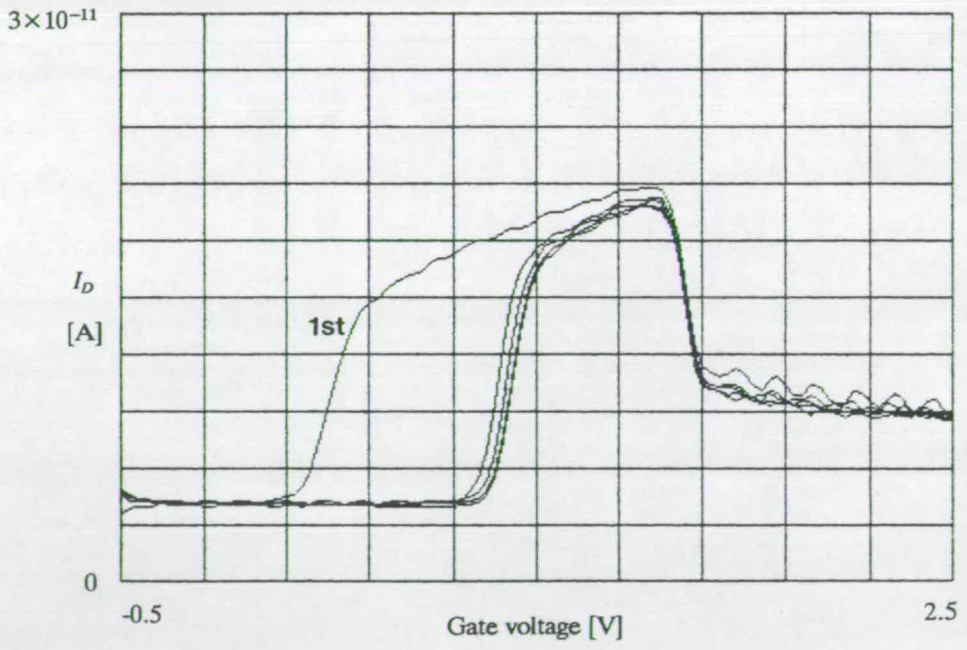


Figure 5.20: Gated diode measurements repeated five times consecutively. $V_{DB} = 1\text{V}$, $V_{GB} = -0.5$ to 2.5V. The source terminal was left floating.

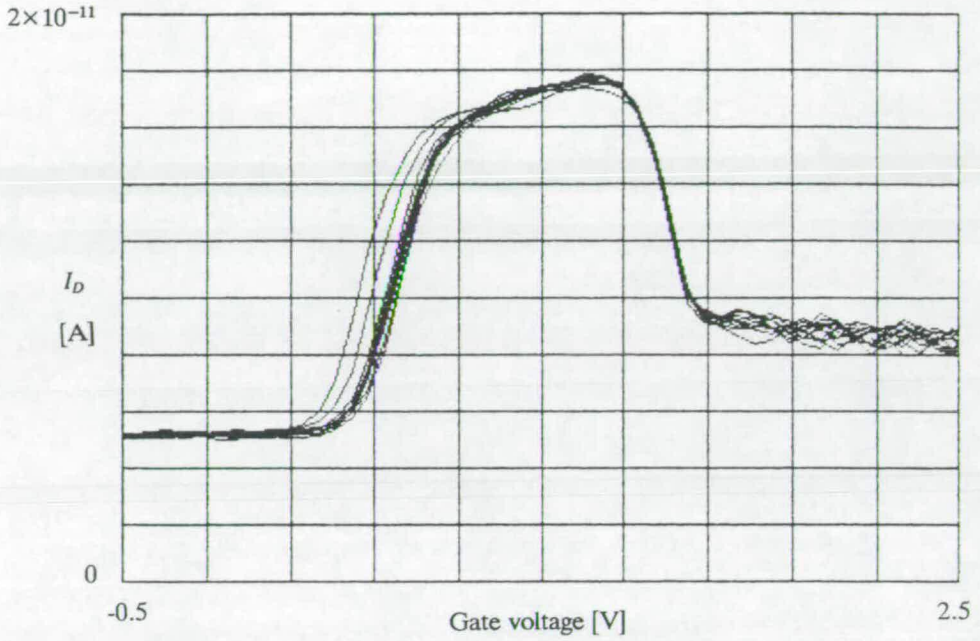


Figure 5.21: Gated diode measurements repeated ten times consecutively. The sample was submitted to both RCA and Karo's clean before gate oxidation. $V_{DB} = 1\text{V}$, $V_{GB} = -0.5$ to 2.5V. The source terminal was left floating.

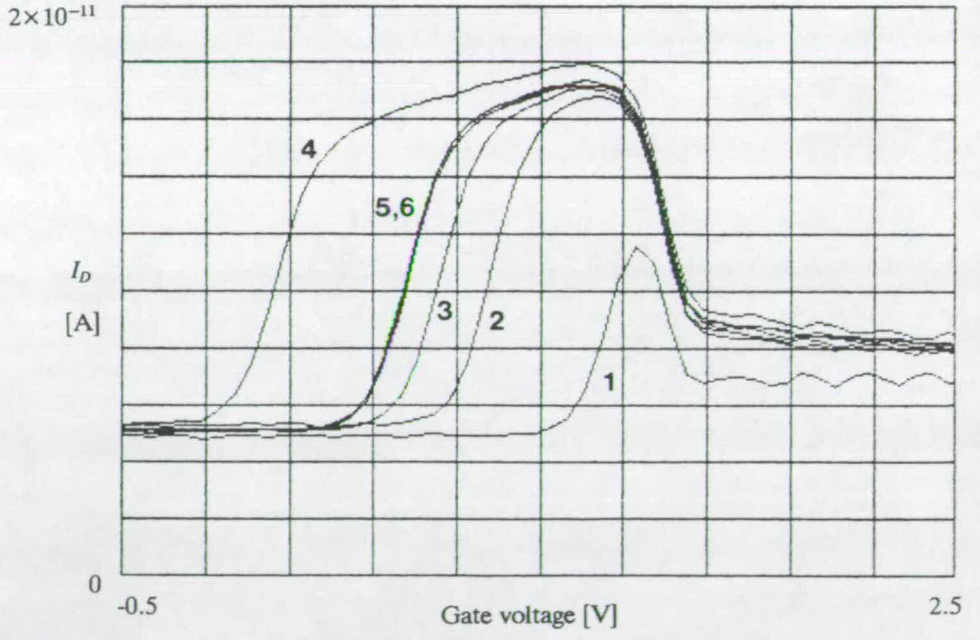


Figure 5.22: Gated diode measurements repeated six times consecutively on the same sample whose initial tests are shown in Figure 5.21. A rest period of 15 hours preceded the final five measurements. The order of measurements is shown in the figure. $V_{DB} = 1\text{V}$, $V_{GB} = -0.5$ to 2.5V. The source terminal was left floating.

this case is that they also include big capacitors for C-V testing. Figure 5.23 shows a typical result of gated diode measurement on a test strip transistor and of a C-V measurement on a capacitor in the same strip. Notice that, although the flatband drift effect is also present in the gated diode curve, there is no evidence of flatband shift in the C-V curve.

After those initial experiments, the following conclusions could be drawn:

a) The problem is not being caused by calibration errors in the measurement system, because it affects the flatband voltage without any change in the threshold voltage;

b) It is not being caused by the test jig. If that was the case, a current leakage would be the observed effect;

c) The voltage shifts observed both after a series of consecutive tests and after a long period rest indicate that there is a presence of some sort of unwanted charge which is not related to the SiO_2 characteristics (i.e., it is not being caused by oxide charges), as concluded from the C-V measurement;

d) There is no correlation between the observed shifts and charge pumping results.

An important fact was then realised: after different voltages were applied in a gated diode test, there was no provision for it to discharge any residual charges still present. Any static charge present on the gate would certainly cause a flatband shift. This hypothesis was easily confirmed by introducing a grounding step (in this case lasting one second) immediately before the test routine. It resulted in curves like in Figure 5.24, which shows a sequence of five consecutive measurements with almost no flatband voltage shift between tests. The problem was, therefore, solved.

5.8. Implementation of the measurement of interface state energy distribution

This requires an adapted charge pumping measurement in which the transition times of the gate pulse are varied, as previously explained in Chapter 4. There are two different ways to vary the transitions: by varying the pulse rise and fall times [17] and by inserting a variable dead band whose voltage level lies between V_{FB} and V_{TH}. [18, 19] The second method is usually easier to implement and the results are easier to process.

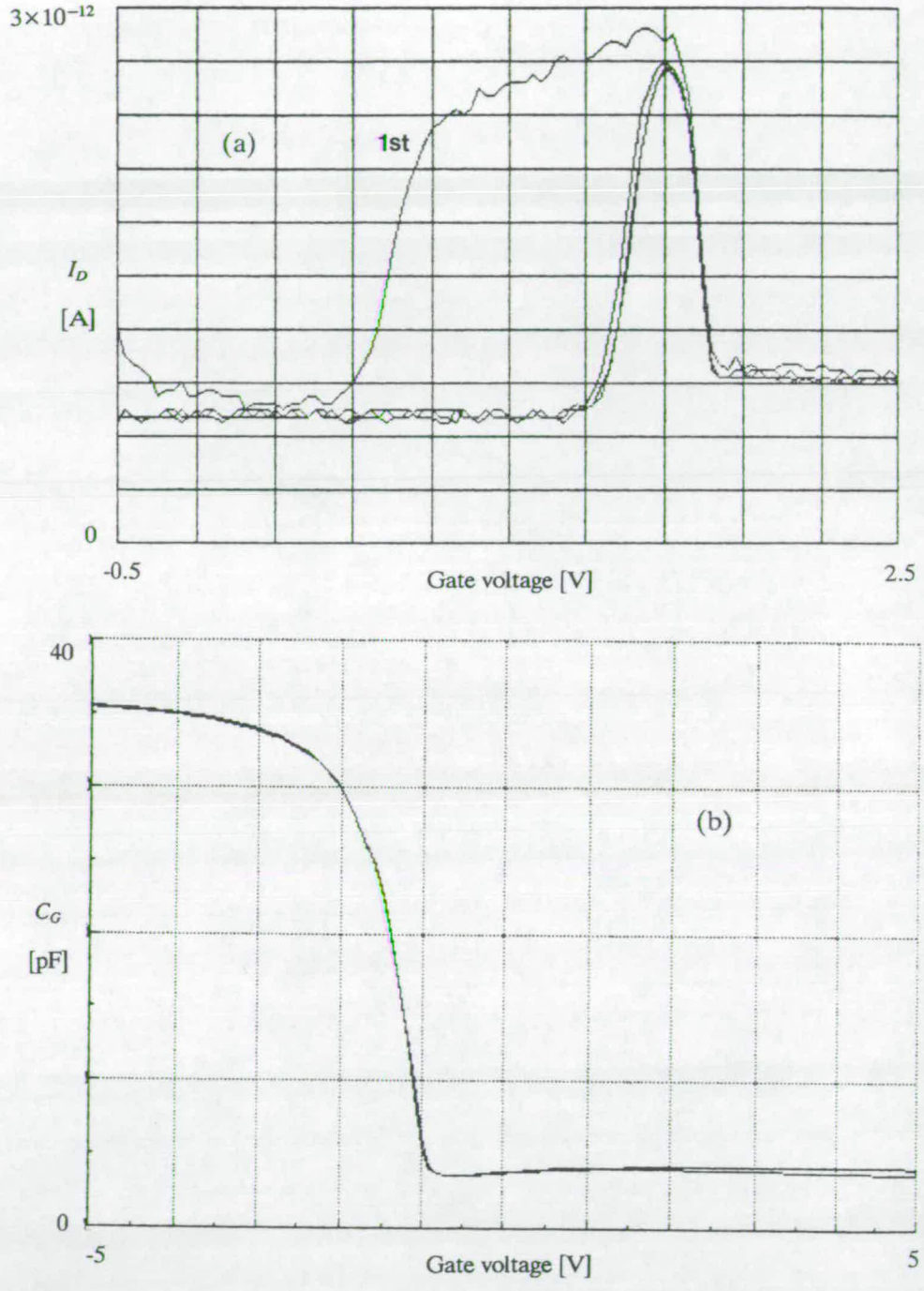


Figure 5.23: a) Gated diode measurements repeated five times consecutively on a transistor belonging to a standard EMF test strip. Notice that the same shift observed in previous samples was also present here. b) Capacitance-voltage result of a capacitor in the same test strip (a hold time of 10 minutes was used between ramping the voltage up and down).

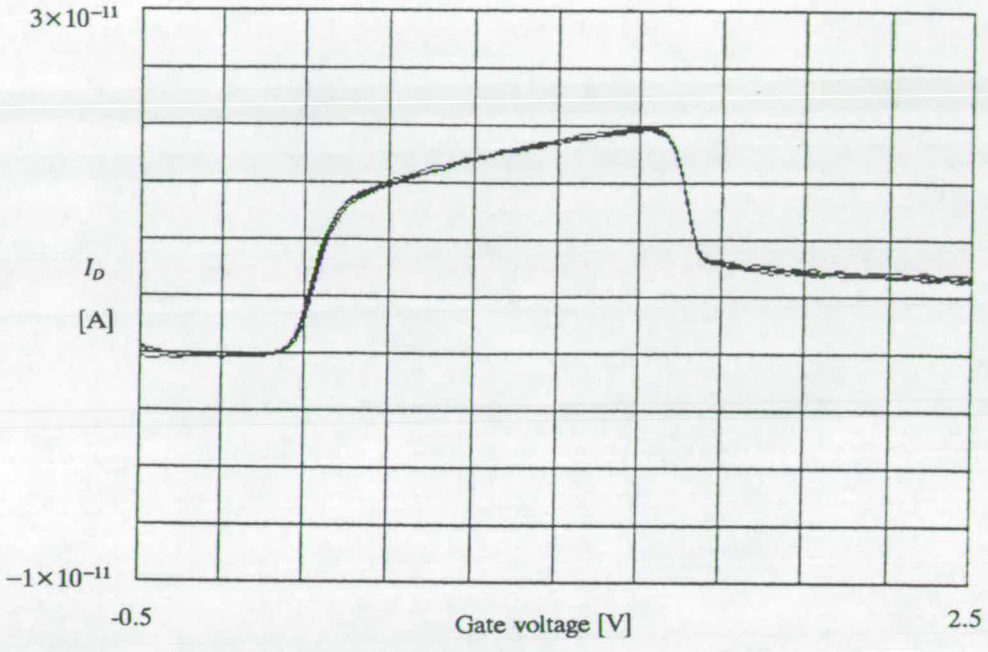


Figure 5.24: Gated diode measurements repeated five times consecutively. A grounding routine was included before each test.

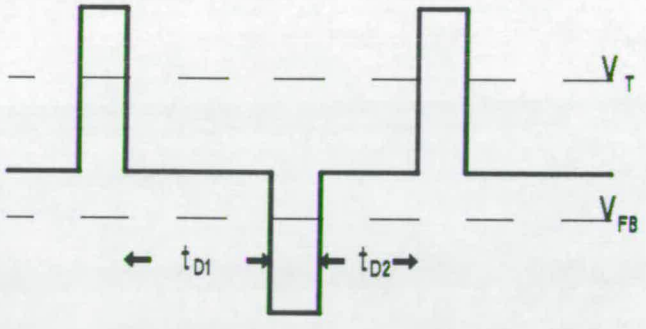


Figure 5.25: Gate pulse waveform used for the adapted charge pumping method used in the determination of the energy distribution of the interface states.

Furthermore, the first method also presents the disadvantage that the interface accumulation does not take place abruptly but starts from the centre of the channel and expands towards the junctions, thus being affected by local variations in interface trap density.[20, 21] The shape of the gate pulse is shown in Figure 5.25. $t_{D,1}$ and $t_{D,2}$ are varied independently, one at a time. The dependence of the energy upon these depletion times is given by [18]

$$\Phi(t_D) = \pm \frac{kT}{q} \ln(\sigma v_{th} n_i t_D) \tag{5.2}$$

where t_D represents either $t_{D,1}$ or $t_{D,2}$ and the sign is positive for the values above midgap and negative below midgap. The relationship between the density of interface states and depletion times is given by

$$D_{it}(t_{it}) = \frac{\Delta i_{CP} T + i_{CP} \Delta t_D}{qLWkT \ln[t_D/(t_D - \Delta t_D)]} \tag{5.3}$$

where T is the total period of the gate signal. A typical example of the resulting curve is shown in Figure 5.26. It has been pointed out by Cilingiroglu [20] that the use of this reverse voltage different from zero implies the adverse presence of a leakage current due to generation in the space charge layers. This new current component would then be more pronounced for energies closer to midgap because the depletion time has to be longer. Therefore he proposes a method which consists of measuring half a transistor in the same fashion as the gated diode method (i.e., leaving either source or drain floating and using the other junction together with different gate biases), this time with a pulsed gate bias. This initial assessment is called pulse interface probing (PIP). The characteristics of the gate pulse are shown in Figure 5.27. When t_D, the time period in which the channel remains under depletion, is varied, a variation in the junction reverse current is obtained, as depicted in Figure 5.28. This curve is used to extract the Preier parameters. According to Preier's trap distribution model, [22, 23] the trap time constant follows the distribution

$$\tau = \tau(0) \exp(x/x_D) \tag{5.4}$$

where x is the distance from the interface into the gate oxide and x_D is a decay constant. The distribution is usually represented as

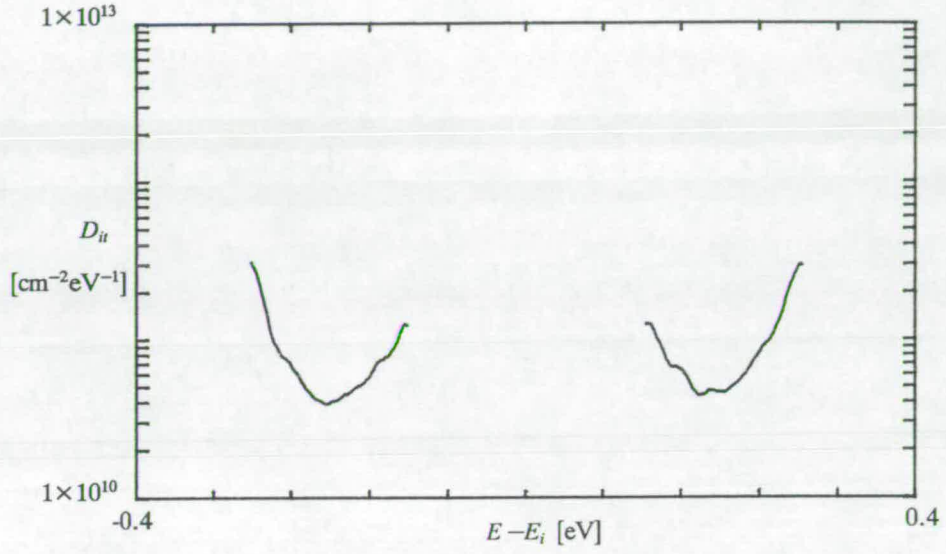


Figure 5.26: Example of energy distribution of interface states in a 1.5μm, non-gapped transistor.

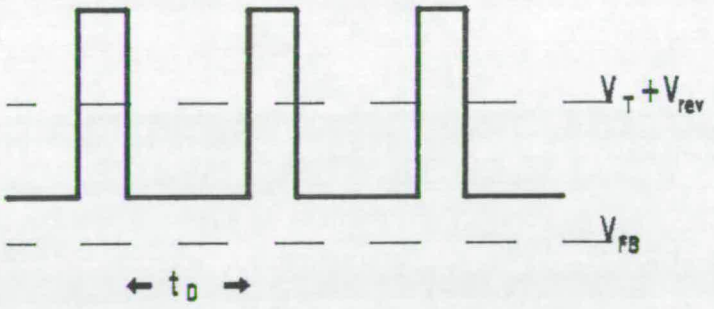


Figure 5.27: Gate pulse waveform used for the pulsed interface probing (PIP) measurements.

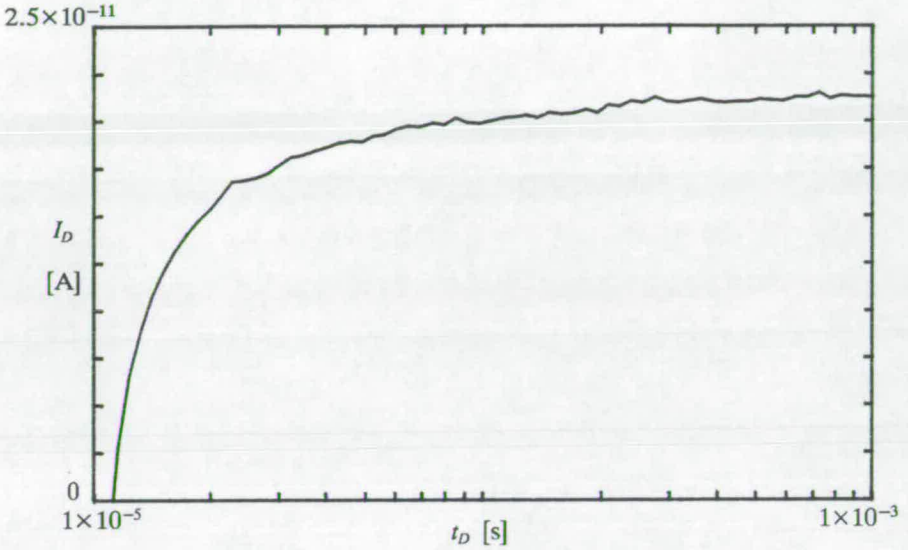


Figure 5.28: Example of a PIP result. $V_{GB,lo} = -3\text{V}$, $V_{GB,hi} = 2\text{V}$, $V_{DB} = 0$. The source terminal was left floating.

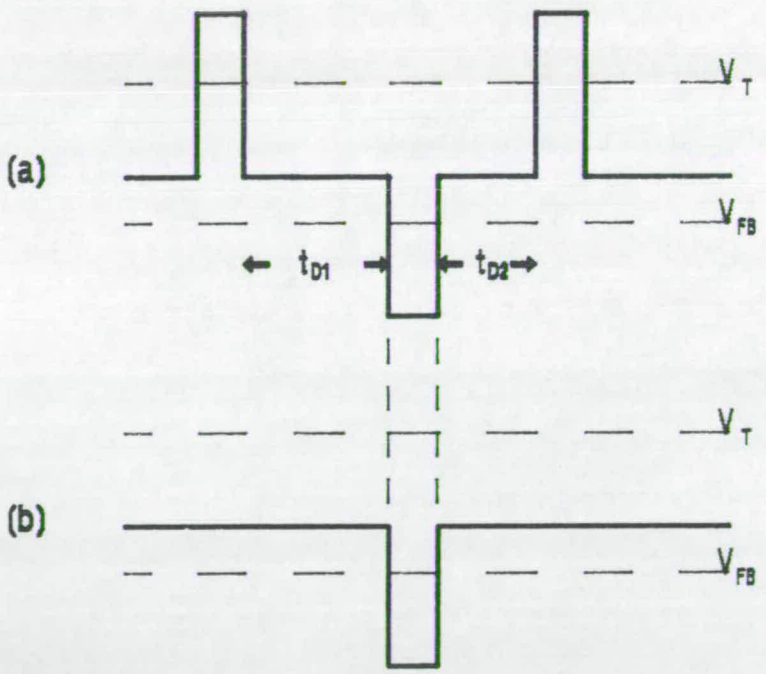


Figure 5.29: Gate pulse wafeforms used to avoid leakage currents at energies near the midgap when using charge pumping for the determination of the energy distribution of D_{it}. The measurement is done with waveform "a" and repeated with waveform "b". The difference between the two charge pumping curves is used to compute D_{it} versus energy.

$$D_{it} = \frac{x_D D_{it}}{\delta\tau} \tag{5.5}$$

and consequently

$$\tau_0 = \frac{\delta\tau(0)}{x_d} \tag{5.5}$$

The parameters we need to extract are τ_0 and δ/x_d. After the curve of Figure 5.28 is obtained, the current resulting from applying a d.c. voltage to the gate ($V_{rev} - V_T < V_{d.c.} < V_{FB}$) is subtracted from the a.c. results. Using

$$\Delta I_R \Big|_{T_D \to \infty} = \frac{D_{it0}}{2} \tag{5.6}$$

and

$$\frac{d\,\Delta I_R}{dT_D}\Bigg|_{T_D \to 0} = \frac{\pi D_{it0}}{2\tau_0} \tag{5.7}$$

we obtain the values of D_{it0} and τ_0. The Preier parameters are then calculated from

$$\frac{d\,\Delta I_R}{dT_D}\Bigg|_{T_D = \tau(0)} = 0.457\,\frac{D_{it0}}{\tau_0} \tag{5.8}$$

and

$$\tau_0 = \frac{\delta\tau(0)}{x_d} \tag{5.9}$$

These parameters are then used to calculate the aperture and peak location of the energy window in the bandgap. Initially a double pulse, charge pumping measurement (in which the pulse has the characteristics of Figure 5.29.a) is employed. Measurements are then repeated with pulses like those depicted in Figure 5.29.b and the obtained results are subtracted from the preceding double pulse measurement. This has the purpose of avoiding the leakage component discussed previously. The density of interface states is then calculated as a function of the aperture previously obtained and plotted against the peak location. However, it was observed that the ΔI_R versus T_D curve (from which the Preier parameters are calculated) was extremely sensitive to the junction reverse bias, which was not discussed in the original literature. [23]

Figure 5.30 depicts this curve for different junction biases. Note that not only is there a current offset amongst different curves, but the slopes are also different. This would inevitably lead to an ambiguous determination of the Preier parameters. The method was, therefore, discarded.

However, the idea of using the pulses shown in Figures 5.29.a and 5.29.b could still be used in order to avoid the leakage component without having to use Preier parameters to calculate the energy values. Therefore the same conventional method used for calculating the energies from the depletion times in the double pulse method (Equation 5.2) was repeated, this time with an additional measurement using the pulse shape depicted in Figure 5.29.b. The current values obtained were subtracted from the current measured previously using the double pulses. The result was then used in the computation of D_{it} (using Equation 5.3). This modified method is compared to the conventional double pulse technique in Figure 5.31. We can see that the procedure in fact increases the extracted density of interface states near the midgap instead of decreasing it. Therefore the method adopted for the determination of the energy distribution of interface states in the present work was the conventional double pulse technique as originally suggested by Hofmann,[18] which uses a single waveform with no further corrections.

5.9. Degradation assessment strategy

In most of the previously published work on hot carrier degradation, the choice of biasing voltages had the objective of maximising the substrate current.[24, 25] This choice certainly implies in a maximised availability of carriers due to avalanche multiplication, but represents only one of the possible conditions which lead to hot carrier generation in IGFETs. In the present work, however, there is an extra aspect to consider. When different transistors are compared,[26] particularly when they have different gate offsets, it is important to keep the electric field distribution as similar as possible, apart from the differences introduced by the various gate positions. This similarity amongst field configurations would hardly be attained if we tried to maximise substrate currents for each individual gate offset. Furthermore, the degradation effects in offset-gate transistors are much more noticeable, even after a short stress routine, as it will be seen later in this Chapter.

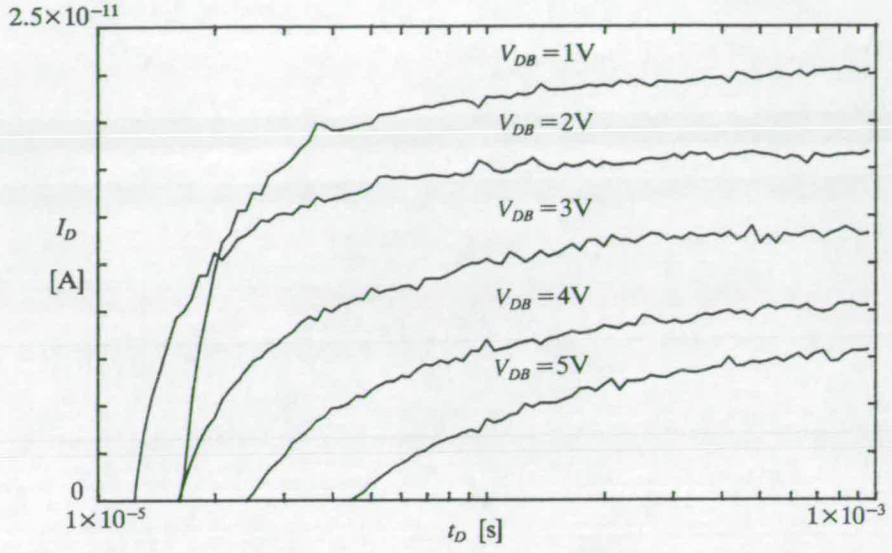


Figure 5.30: PIP measurements repeated for different reverse biases. $V_{GB,lo} = -3\text{V}$ and $V_{GB,hi} = 2\text{V}$. The source terminal was left floating.

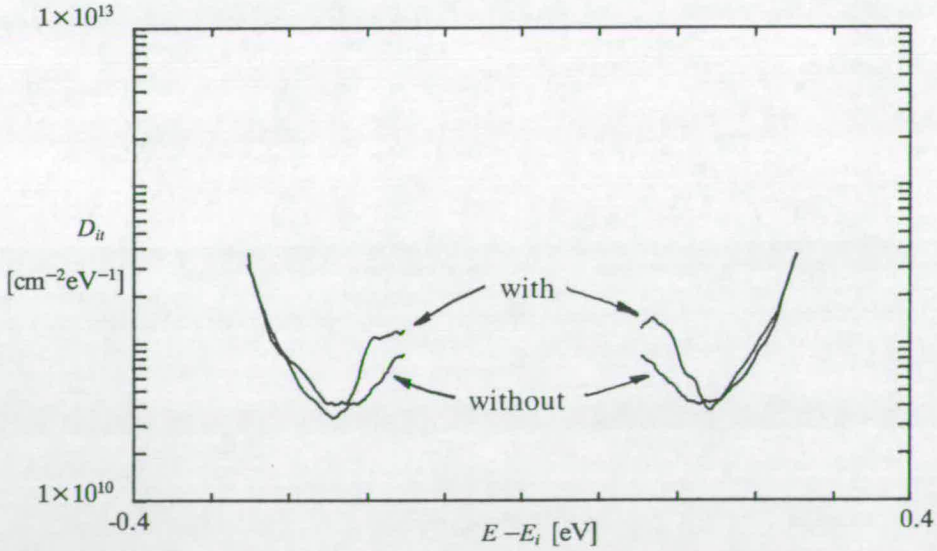


Figure 5.31: Comparison between the energy distribution of interface states computed with and without the extra measurement employing the waveform depicted in Figure 5.29.b

Three different degradation biases were used:

1) $V_{DS}=5V$, $V_{GS}=5V$. Under this condition, only hot electrons are generated, as previously discussed in Chapter 3.

2) $V_{DS}=5V$, $V_{GS}=1V$. This biasing condition favours hot hole injection and was chosen to verify the presence of such injection under different gate offset conditions.

3) $V_{DS}=5V$, V_{GS} pulsed between 0 and 5 V with a frequency of 100 kHz.

During the degradation cycles the currents were continuously monitored by the computer in order to detect any possible accidental alterations to the initial conditions (bad contacts, vibration, accidental opening of the prober cage lid, etc.). Since the same HP4062B parametric tester which was used for all the measurements was also employed for the degradation cycles themselves, this continuous measurement also ensured that the effective current compliance (which in the case of the HP4062B depends on the current requirements during testing) was effectively being maintained.

5.10. Experimental results

Several different groups of transistors with both 1.0 and 1.5 μm channel length were submitted to the degradation tests. The following procedure was used in most of the cases:

1) Transistor gap assessment using subthreshold current measurement ($V_{DS}=1\ V$, $V_{GS}=0$ to 5 V) in both direct and reverse configurations;

2) Charge pumping current measurement with source, drain and substrate grounded and gate pulses oscillating with an amplitude of 5 V at 250 kHz and a varying offset so that the base level goes from -8 V to 2 V;

3) Gated diode current measurement with source floating, $V_{DS}=1\ V$, V_{GS} varying from -0.5 to 2.5 V and the substrate grounded;

4) 1000 second degradation cycle with $V_{DS}=V_{GS}=5\ V$ and source and substrate grounded;

5) Gated diode current measurement with source floating, $V_{DS} = 1\ V$, V_{GS} varying from -0.5 to 2.5 V and substrate grounded;

6) Charge pumping current measurement with source, drain and substrate grounded and gate pulses oscillating with an amplitude of 5 V at 250 kHz and a varying offset so that the base level goes from -8 V to 2 V;

7) 10000 second degradation cycle with $V_{DS} = V_{GS} = 5\ V$ and source and substrate grounded;

8) Gated diode current measurement with source floating, $V_{DS} = 1\ V$, V_{GS} varying from -0.5 to 2.5 V and the substrate grounded;

9) Charge pumping current measurement with source, drain and substrate grounded and gate pulses oscillating with an amplitude of 5 V at 250 kHz and a varying offset so that the base level goes from -8 V to 2 V.

Figure 5.32 shows charge pumping and gated diode results for 1.5 μm transistors whose gate offset range from 0 to 0.25 μm before degradation and after the 1000 and 10000 second cycles. The first important characteristic to be observed in this series of measurements is the clear presence of a degraded component shifted towards more positive values with respect to the undegraded curve, which could be indicative of the generation of acceptor-type interface traps as discussed previously in Chapter 3. The height of this secondary curve changes according to the gate/drain gap and peaks for a gap of about 0.2 μm. This tendency was observed for all the transistor groups tested. On the other hand, a consistently increasing flatband shift can be noticed as the gap is increased.

The peak of the charge pumping curve is directly proportional to the average density of interface states, $\overline{D}_{it}$. The absolute value of this average density can be determined by measuring the charge pumping current using a fixed source/drain bias, fixed low and high values and triangular gate pulses with variable frequency and plotting the charge recombined per cycle Q_{SS} given by

$$Q_{SS} = \frac{I_{CP}}{f} \tag{5.10}$$

versus the pulse frequency. Figure 5.33 shows the results of such measurements for a transistor in the same wafer as the group whose results are displayed in Figure 5.32. The

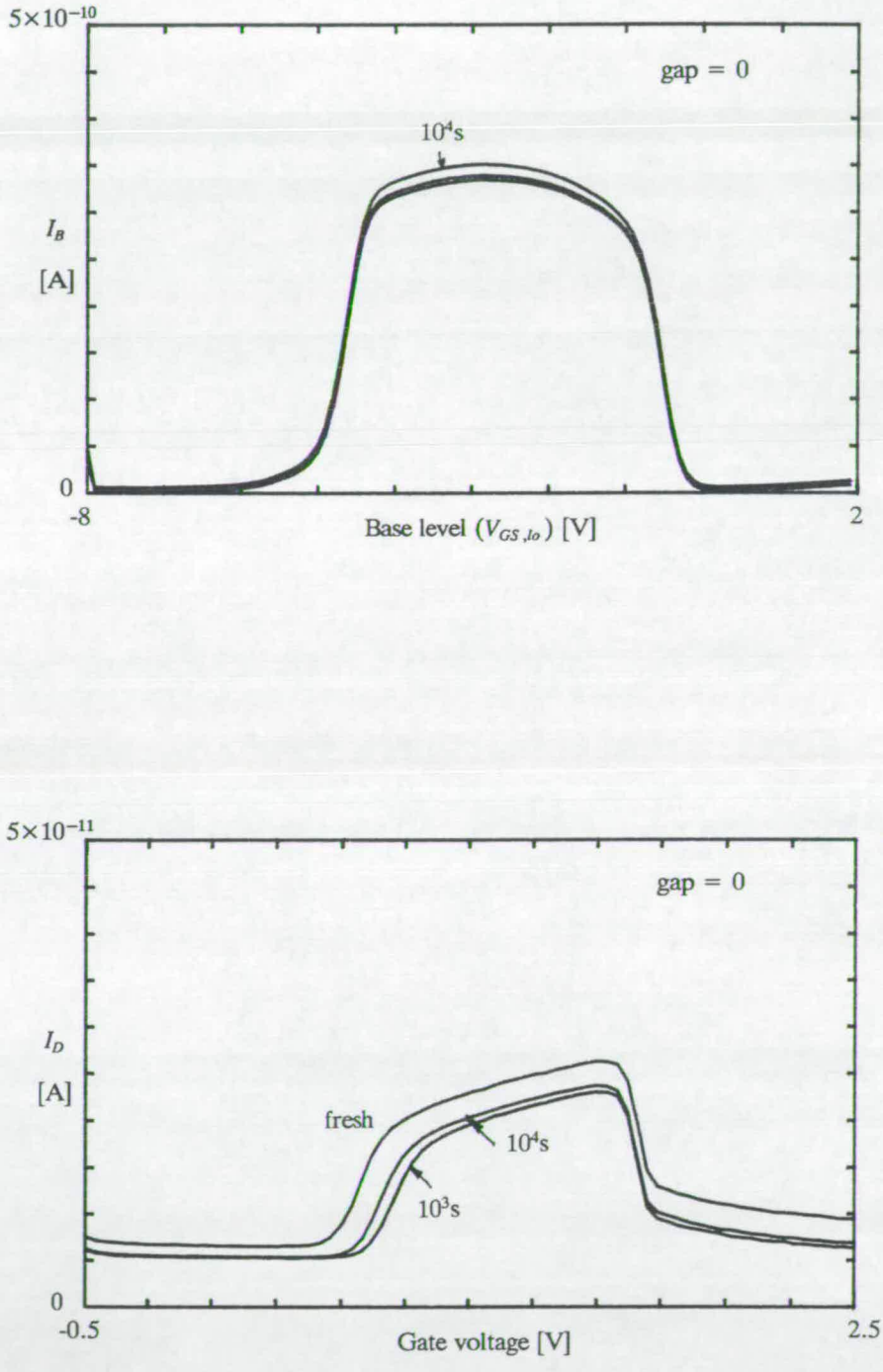
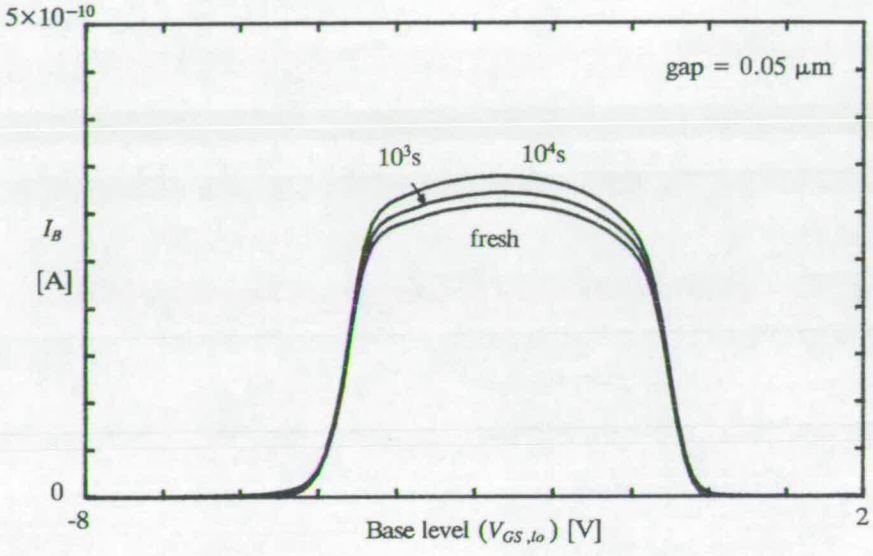


Figure 5.32: Charge pumping and gated diode results of 1.5 μm transistors with gaps ranging from 0 to 0.25 μm. Each graph depicts measurements before and after 1000 and 10000-second stress cycles. Stress was performed with $V_{GS} = V_{DS} = 5\text{V}$ and $V_{SB} = 0$.



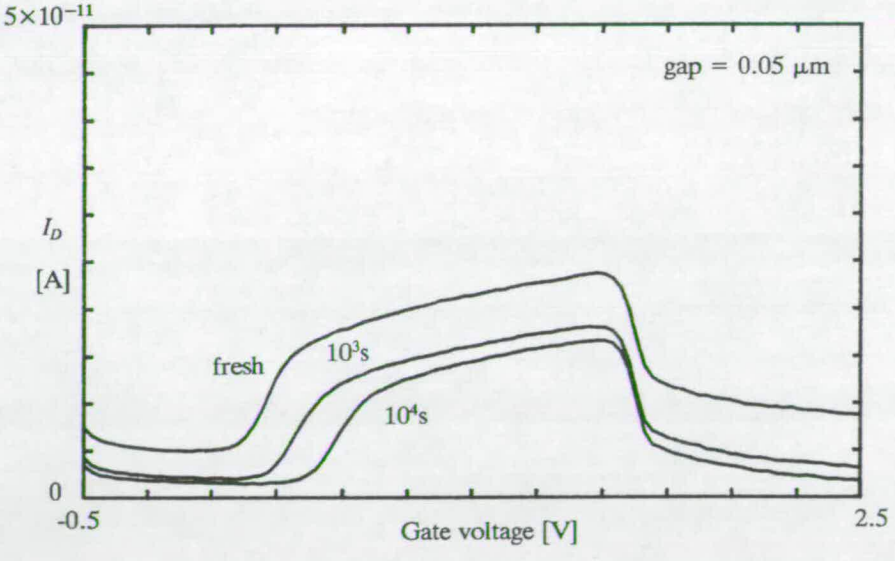
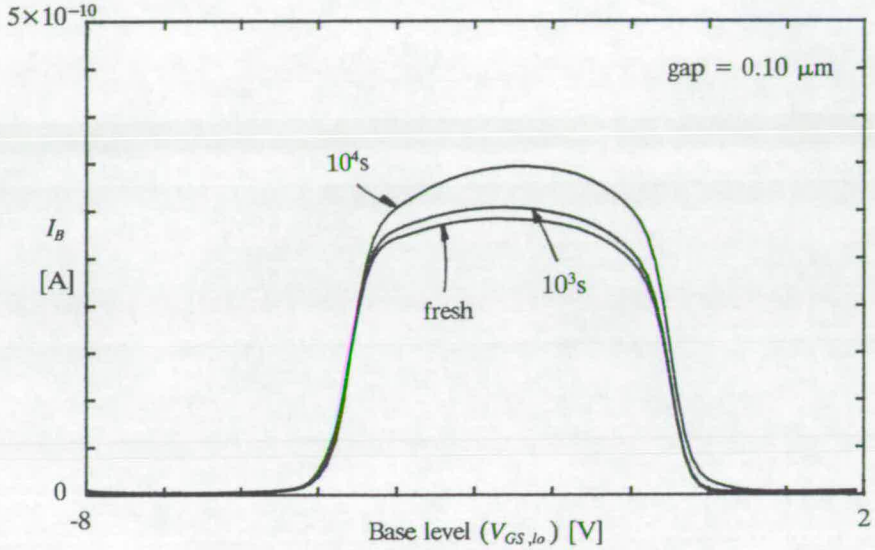


Figure 5.32 (cont'd)



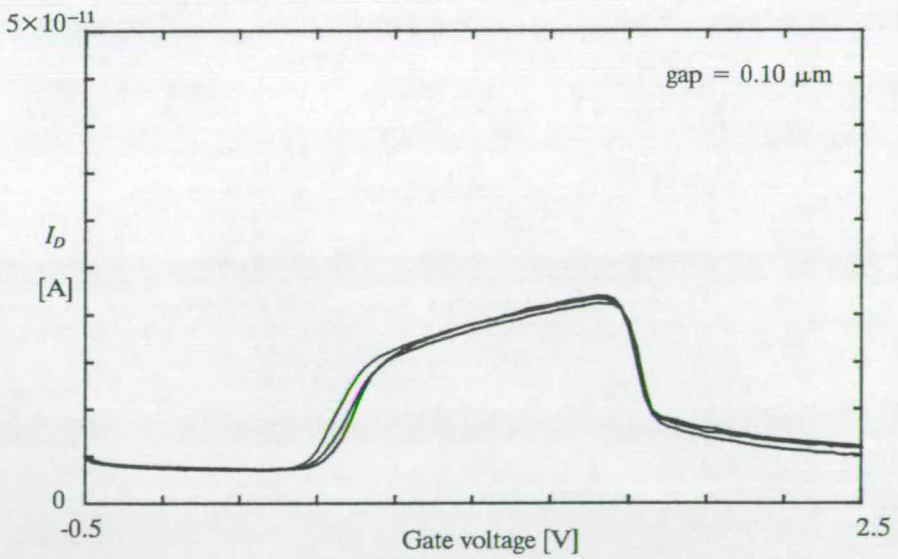
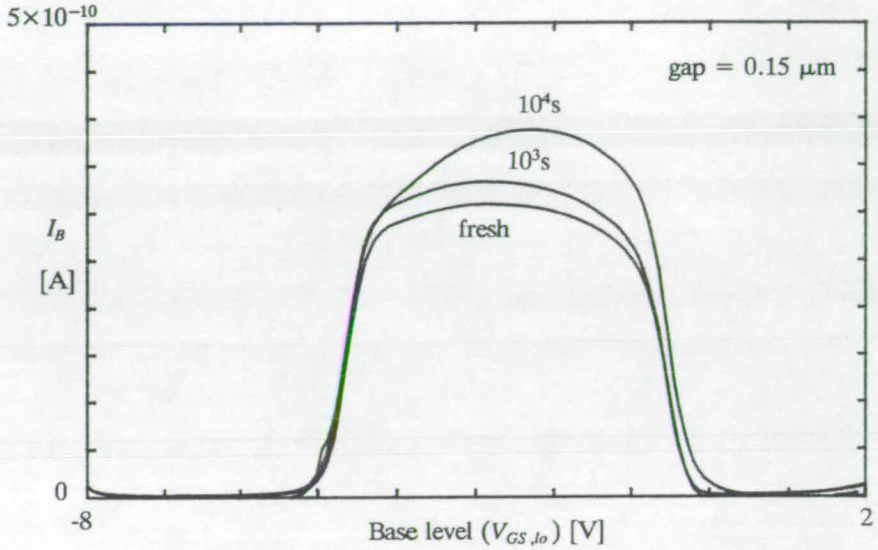


Figure 5.32 (cont'd)



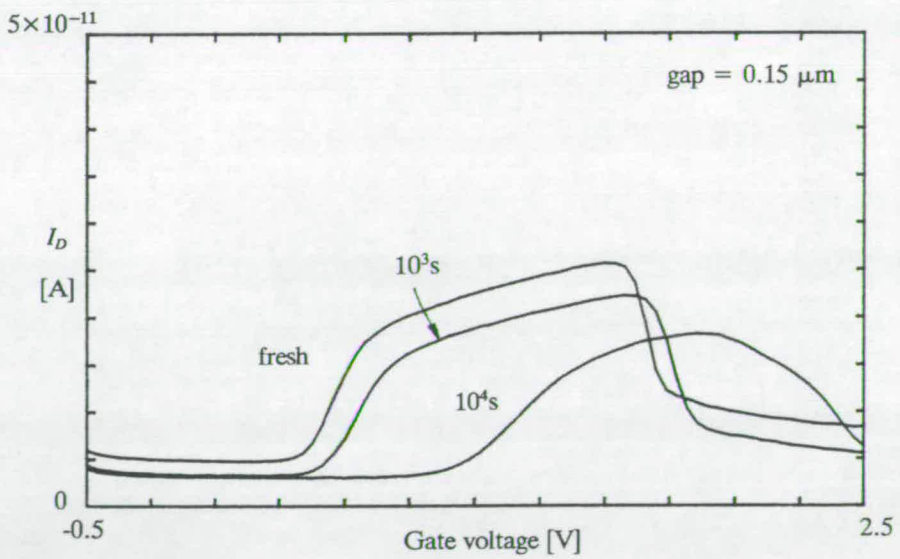
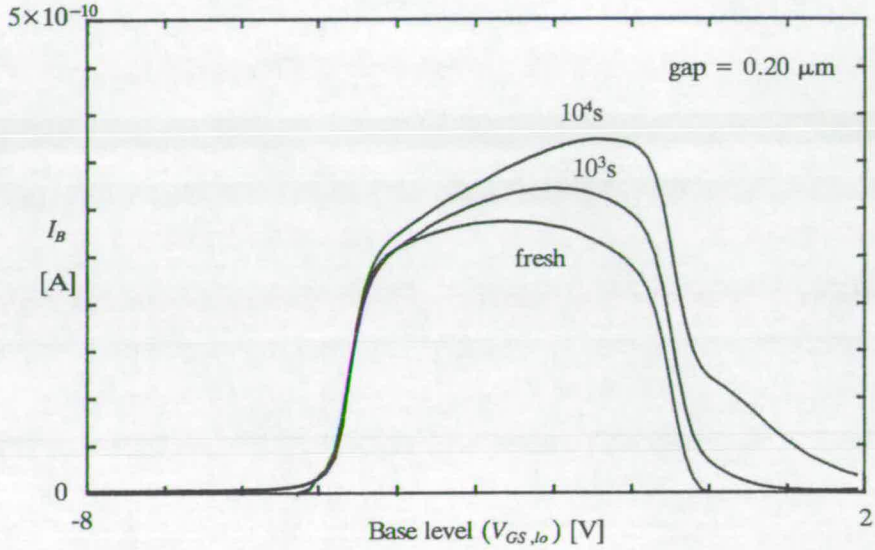


Figure 5.32 (cont'd)



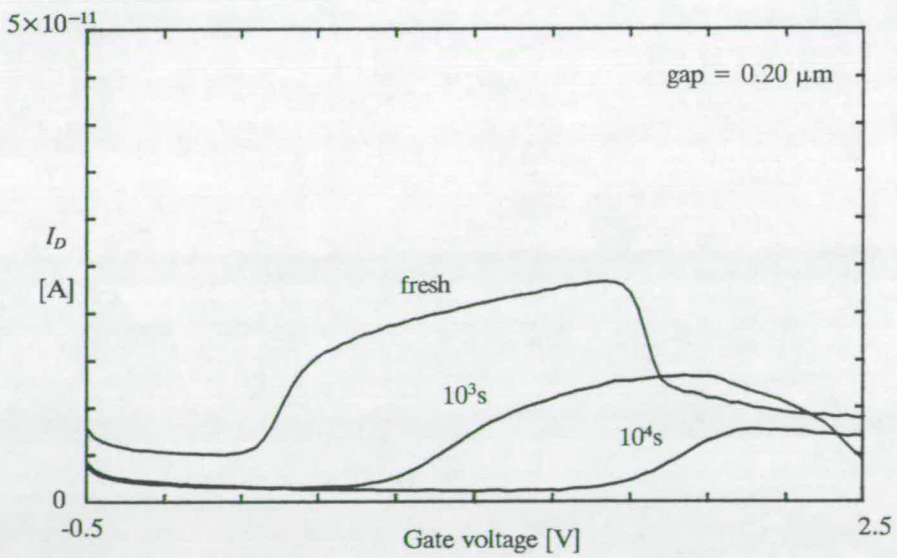
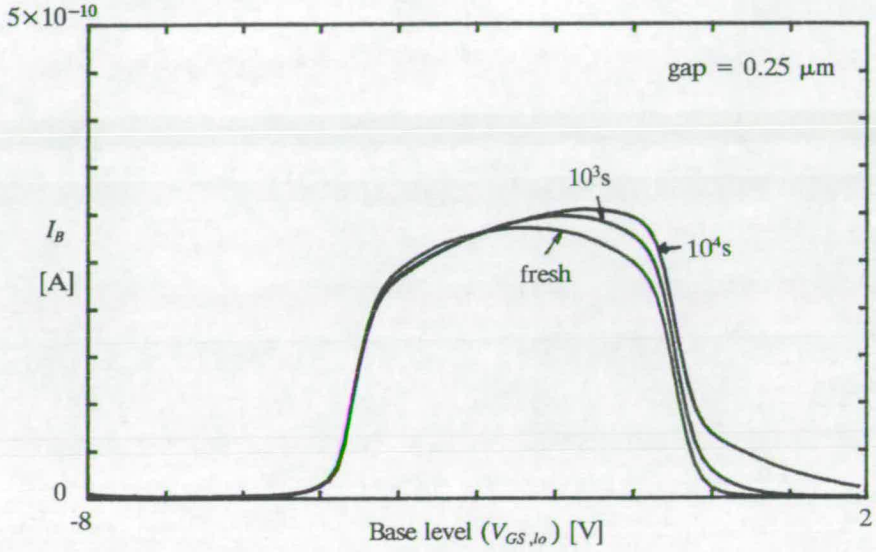


Figure 5.32 (cont'd)



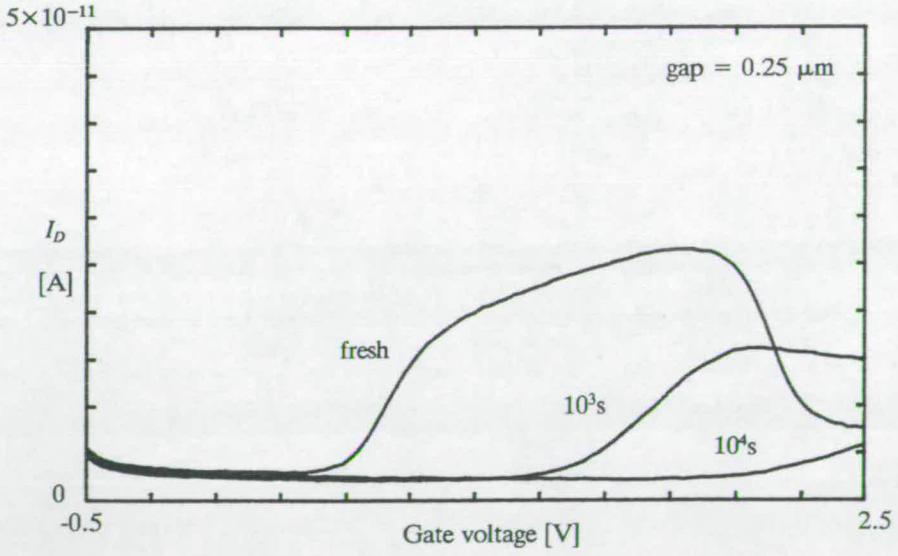


Figure 5.32 (cont'd)

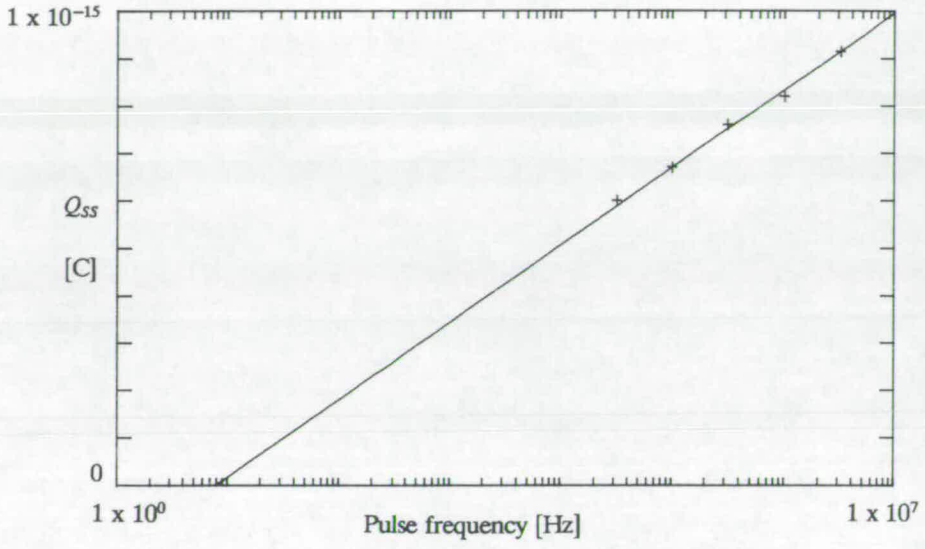


Figure 5.33: Recombined charge per cycle versus pulse frequency. $V_{GS,lo} = -3\text{V}$, $V_{GS,hi} = 2\text{V}$, $V_{DS} = V_S = V_B = 0$.

slope gives [17]

$$\frac{dQ_{SS}}{d\log f} = \frac{2qkT\bar{D}_{it}}{\log e} A_G \tag{5.11}$$

where A_G is the gate area. The x-intercept leads to the geometrical mean of the capture cross sections, namely

$$\sqrt{\sigma_p \sigma_n} = \frac{1}{v_{th} n_i} \frac{|\Delta V_G|}{|V_{FB} - V_T|} \frac{f_0}{\sqrt{\alpha(1-\alpha)}} \tag{5.12}$$

where α is the ratio between the rise time and the pulse period. Like in [18] it was observed that these extracted values, particularly the average value of the density of interface states, remains steady over different sites within a wafer. Therefore the measurement of $\bar{D}_{it}$ on a few different sites is usually enough to establish the correlation between the charge pumping curve peaks and the average density of interface states.

The flatband voltage shifts observed through gated diode measurements provide an almost direct way to quantify the presence of trapped charge. This is given by [27]

$$\Delta V_{FB} = -\frac{Q}{C_{ox}} \tag{5.13}$$

$\Delta\bar{D}_{it}$ and trapped charge results were gathered from different groups of transistors, resulting in the plots of Figures 5.34 and 5.35, respectively. They show more clearly the peak corresponding to a gap of 0.20 μm and the continuous increase in trapped interface charge. These results will be further discussed in Chapter 6.

The same test/degradation/test procedure was also run on 1.0 μm transistors. The results are shown in Figure 5.36. It has been observed that the level of degradation is considerably higher than for 1.5 μm transistors and that the peak in the generation of interface states as a function of the gap size occurs at 0.1 μm instead of 0.2 μm.

In previous observations of hot hole degradation, the confirmation that holes were indeed responsible for the alteration in the shape of the charge pumping curve was obtained by performing a short hot electron injection after the main degradation/assessment routine.[16] The holes initially trapped would then be neutralised by the injected hot electrons. This process, however, cannot be easily reversed for the case of hot electron injection, because of the difference in capture cross sections between

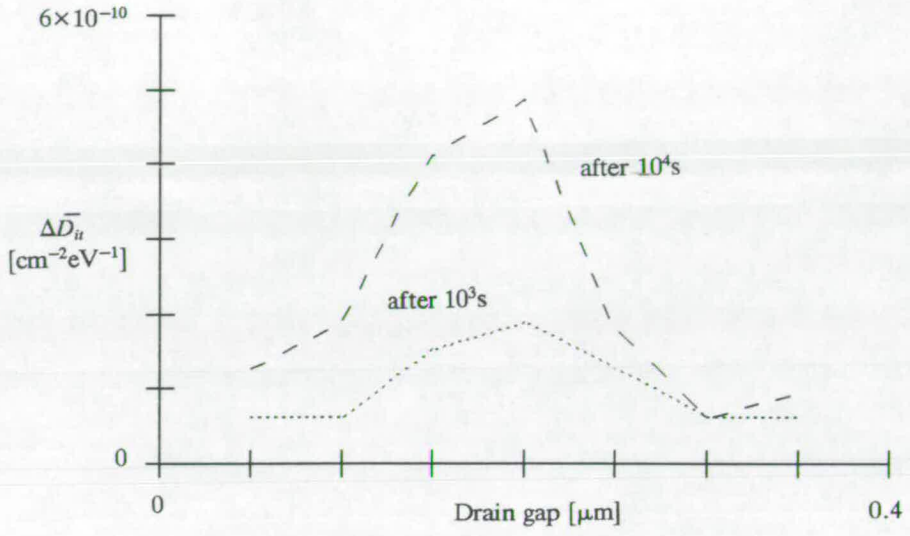


Figure 5.34: $\Delta\bar{D}_{it}$ versus drain gap. $\Delta\bar{D}_{it}$ was averaged over several different groups of transistors.

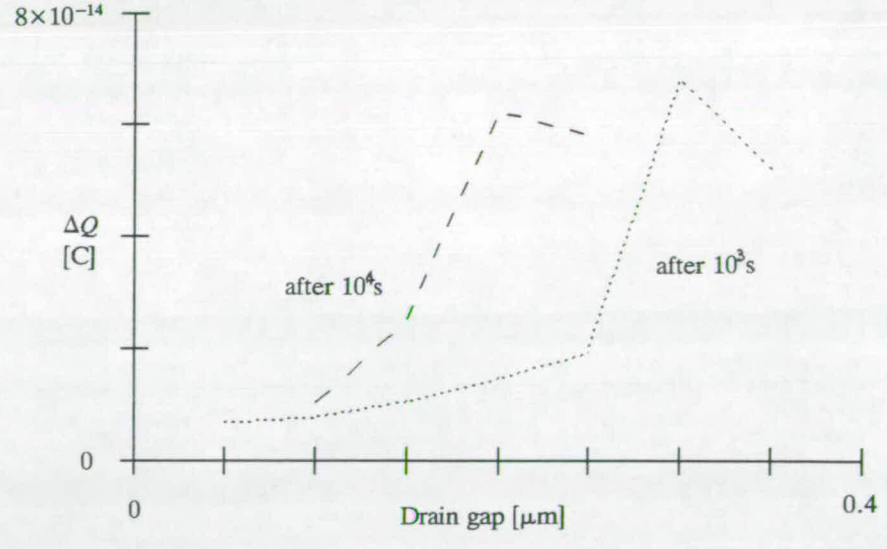
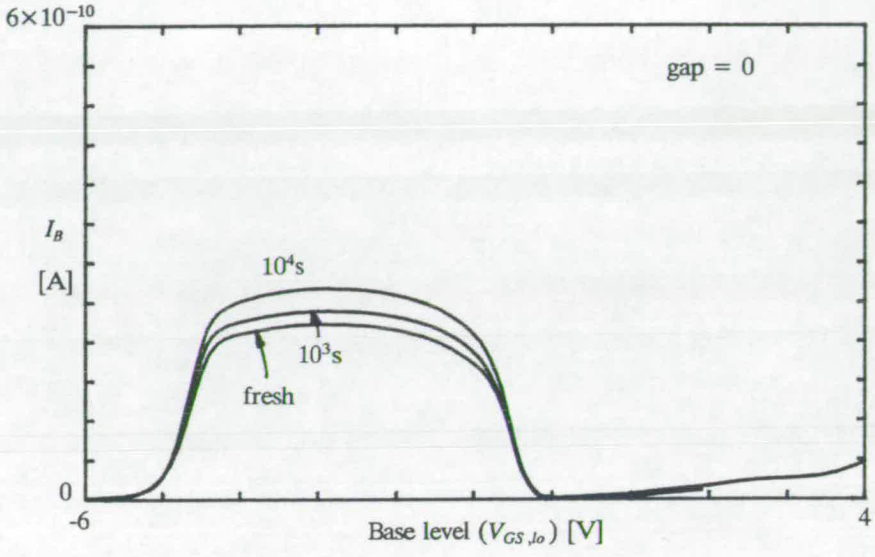


Figure 5.35: ΔQ versus drain gap. ΔQ was averaged over several different groups of transistors.



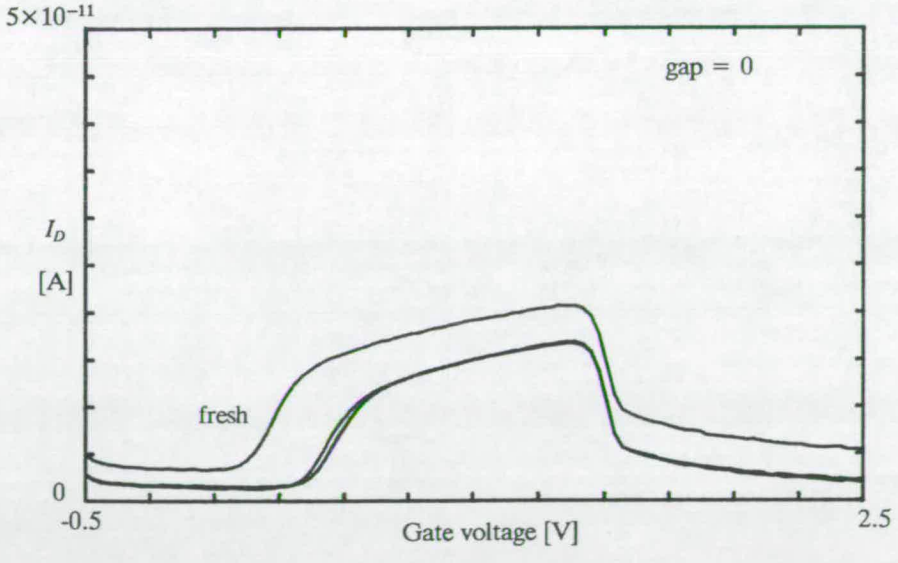
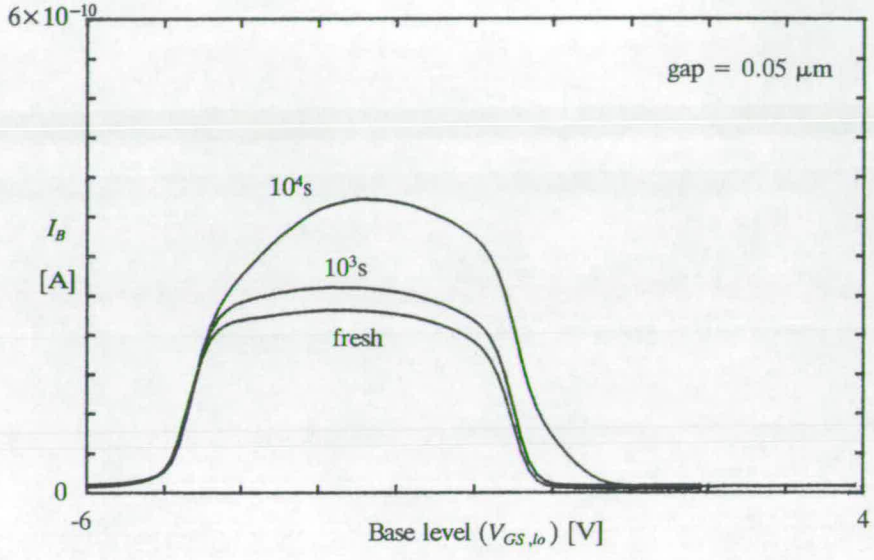


Figure 5.36: Charge pumping and gated diode results of 1.0 μm transistors with gaps ranging from 0 to 0.25 μm. Each graph depicts measurements before and after 1000 and 10000-second stress cycles. Stress was performed with $V_{GS} = V_{DS} = 5\text{V}$ and $V_{SB} = 0$.



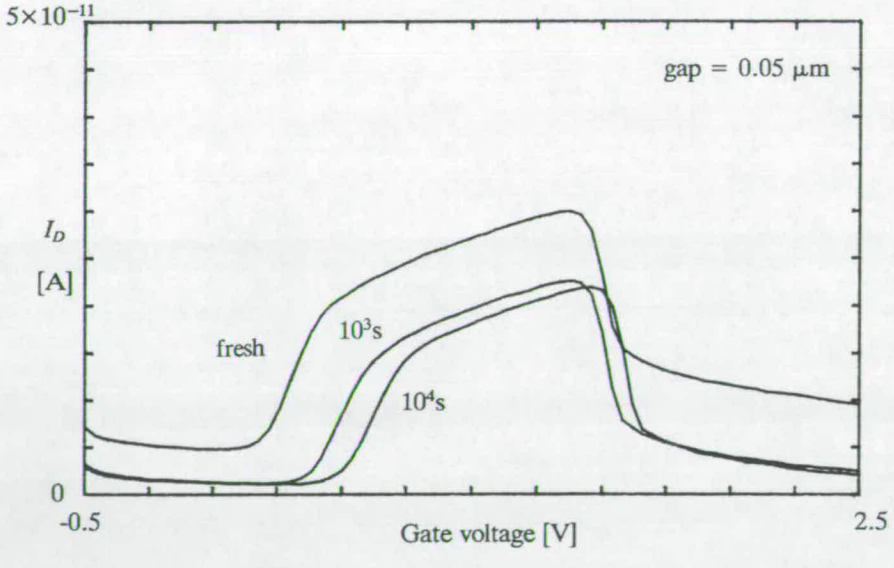
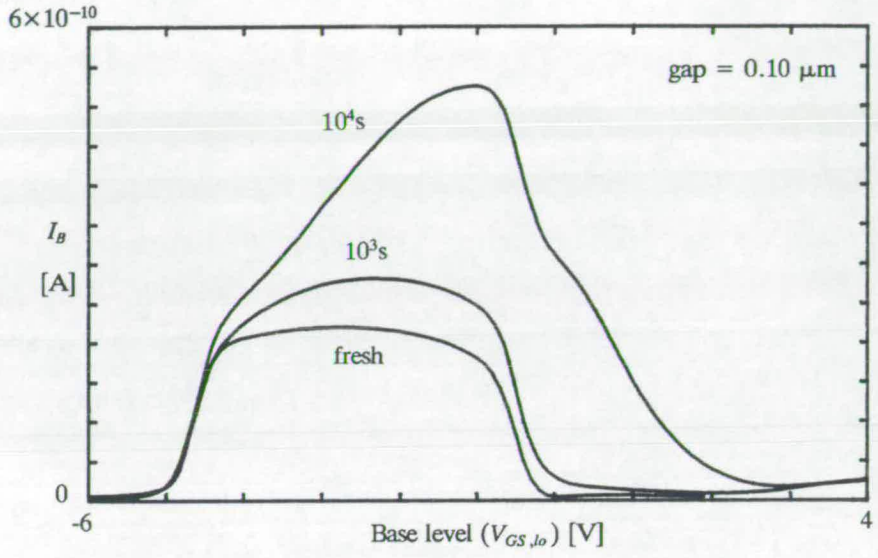


Figure 5.36 (cont'd)



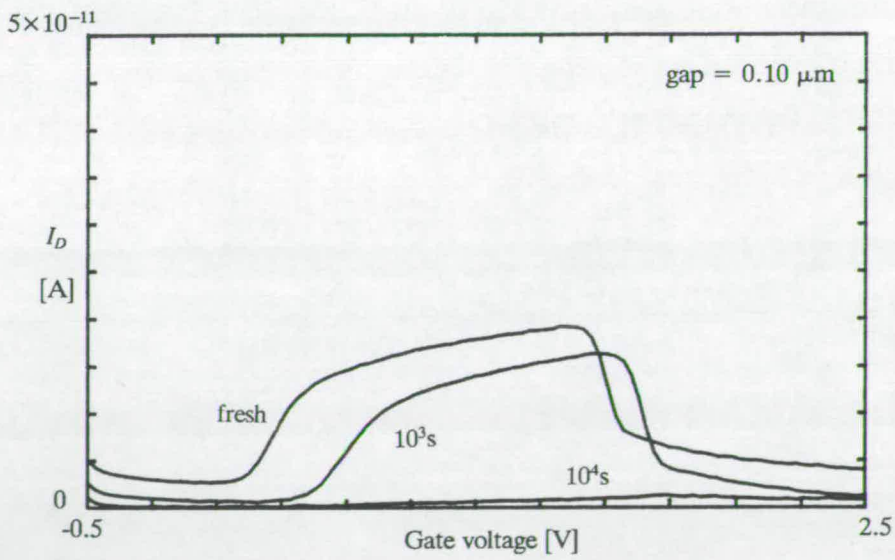
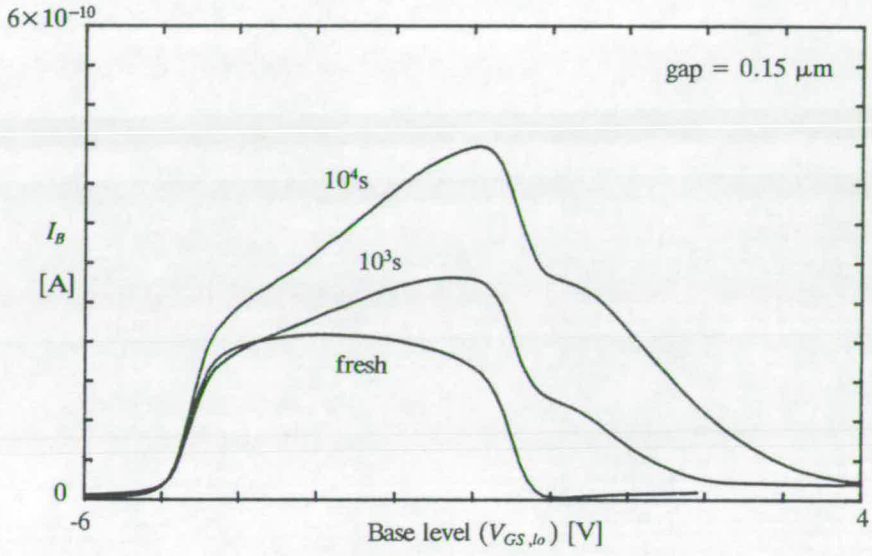


Figure 5.36 (cont'd)



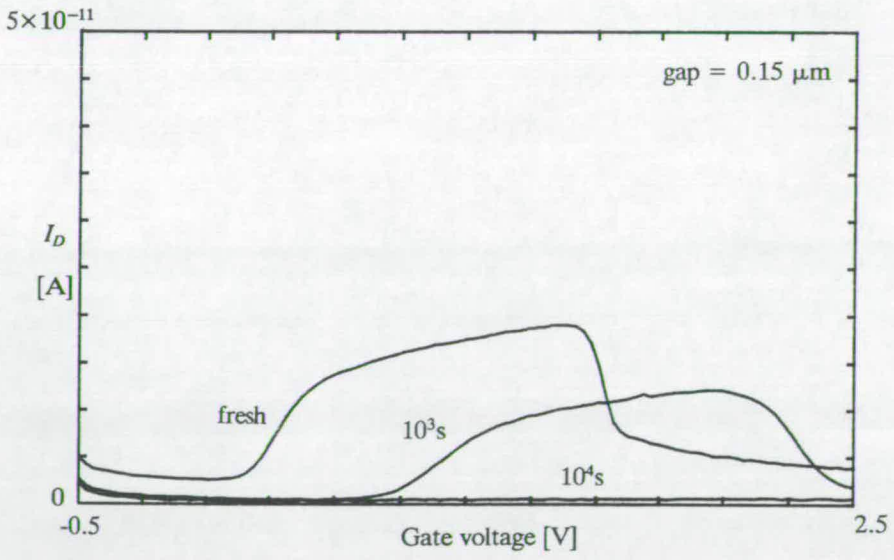
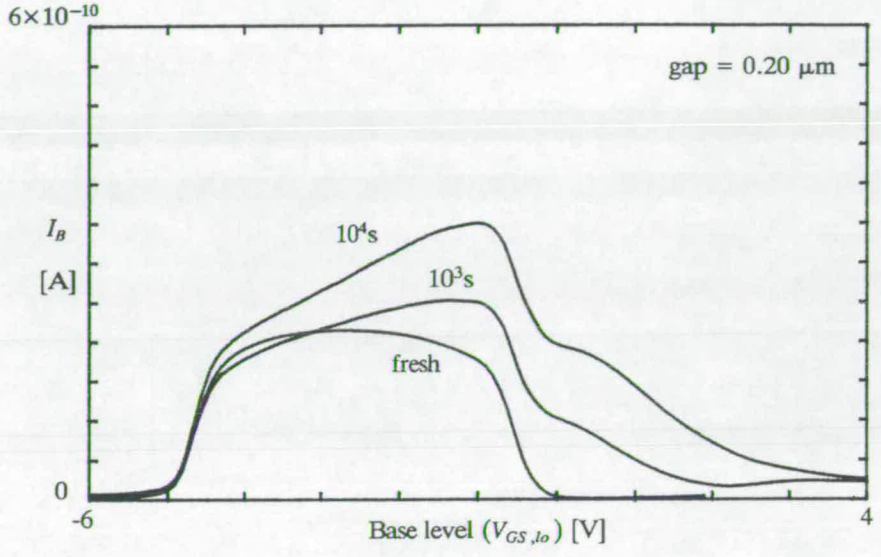


Figure 5.36 (cont'd)



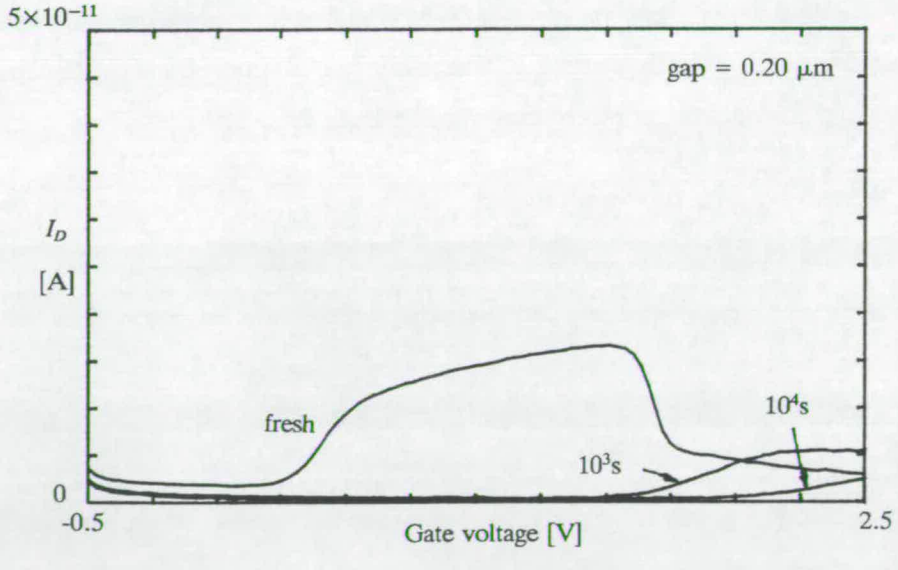
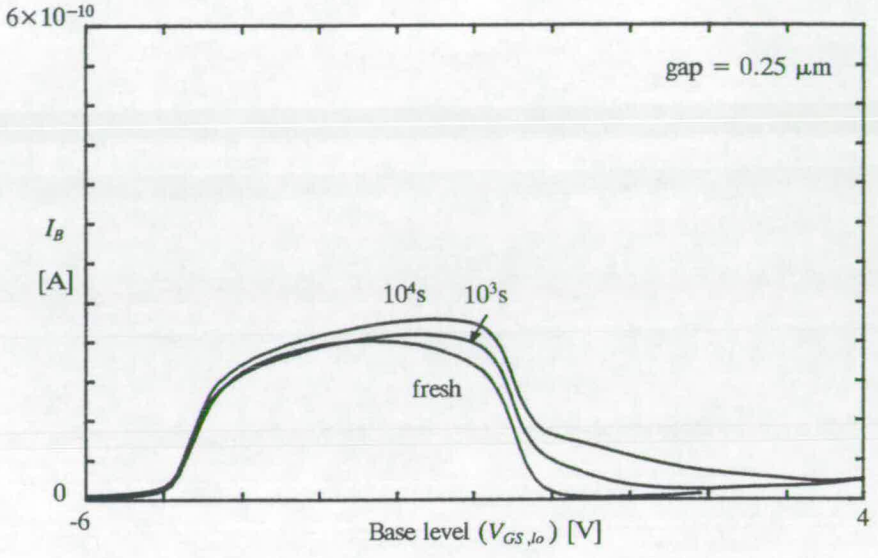


Figure 5.36 (cont'd)



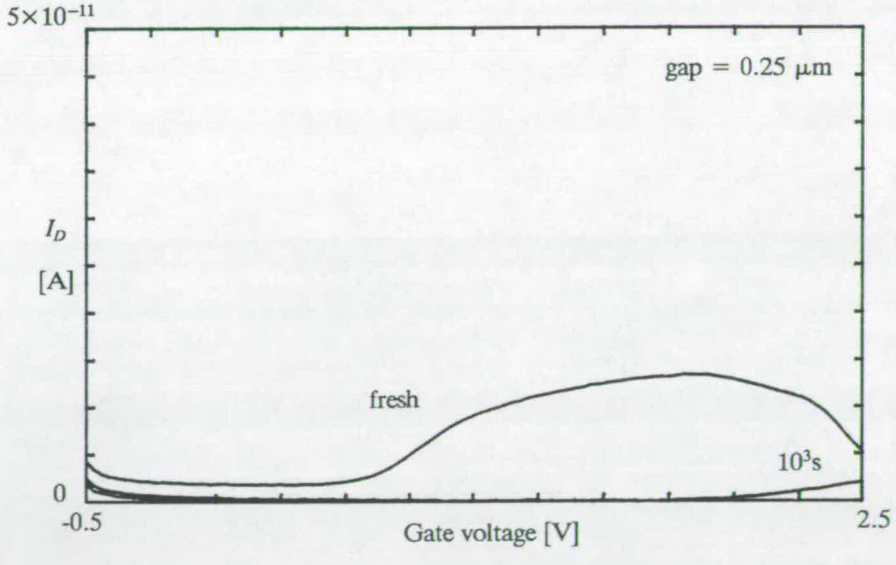


Figure 5.36 (cont'd)

the two different types of carriers. However, some of the injected electrons might have been trapped deeper, therefore with a correspondingly longer time constant. Their detrapping will therefore take place with a higher time constant. Therefore such a confirmation of the presence of electrons trapped as the result of injection consists of simply leaving the samples "resting" for a long period after the degradation assessment and testing them again. This was indeed done for a period of 72 hours and the obtained results are shown in Figure 5.37. As expected, the extended tail (characteristic of electron injection) is considerably reduced after the rest period. However, the peak of the curve remains unchanged, showing that the degradation process indeed created acceptor-type interface traps which remained unchanged during the "self-healing" process.

The next step was to uniquely identify the degradation process as being caused by hot carriers, therefore ruling out eventual concurrent processes such as tunneling. Apart from hot carrier injection, the most likely mechanism for injection into SiO_2 under such circumstances (thin SiO_2 gate insulator, high electric field) is Fowler-Nordheim tunneling, which is the tunneling through the upper triangular portion of the SiO_2 barrier (see Figure 3.1). The current which results from such a tunneling process is given by [28]

$$J = C_1 E^2 \exp(-C_2/E) \tag{5.14}$$

where

$$C_1 = 1.54\times10^{-6} s/\Phi_B \tag{5.15}$$

$$C_2 = 6.83\times10^{7}\Phi_B^{3/2}(m^*/m)^{1/2} \tag{5.16}$$

It is a known fact (see Chapter 2) that at low temperatures there is less lattice vibration and therefore the probability of a hot carrier losing energy through phonon collisions is considerably reduced. Therefore there is an increase in the density of carriers capable of surmounting the Si/SiO_2 barrier and the hot-carrier-related degradation is likely to increase. From the equations for Fowler-Nordheim tunneling given above we can see that it consists of an almost temperature-independent process, in contrast with the low-temperature enhancement observed in hot carrier generation. Therefore, by performing the stress routines at low enough temperatures so that the lattice vibration can be considerably reduced, it is possible to confirm the predominance of hot carrier effects. The

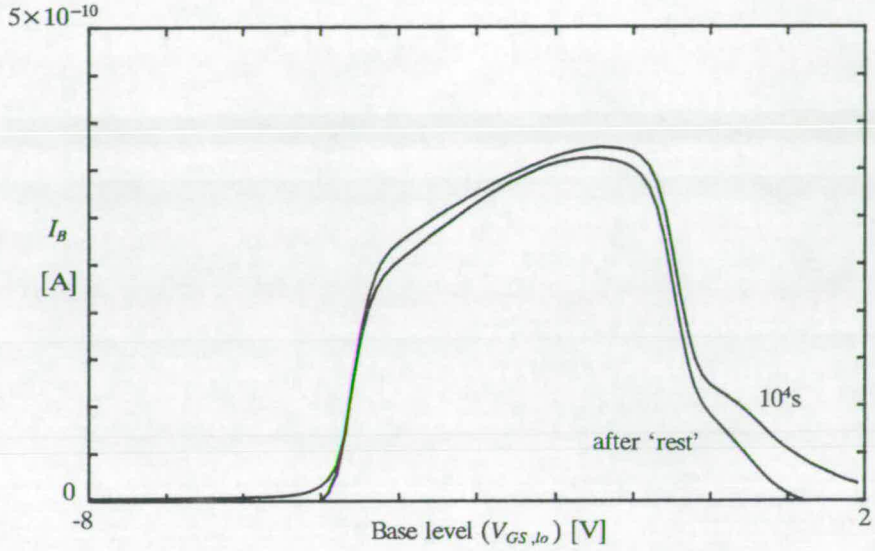


Figure 5.37: Charge pumping results of a 0.2 μm transistor after a 10000 second degradation cycle and after a subsequent 72 hour rest period.

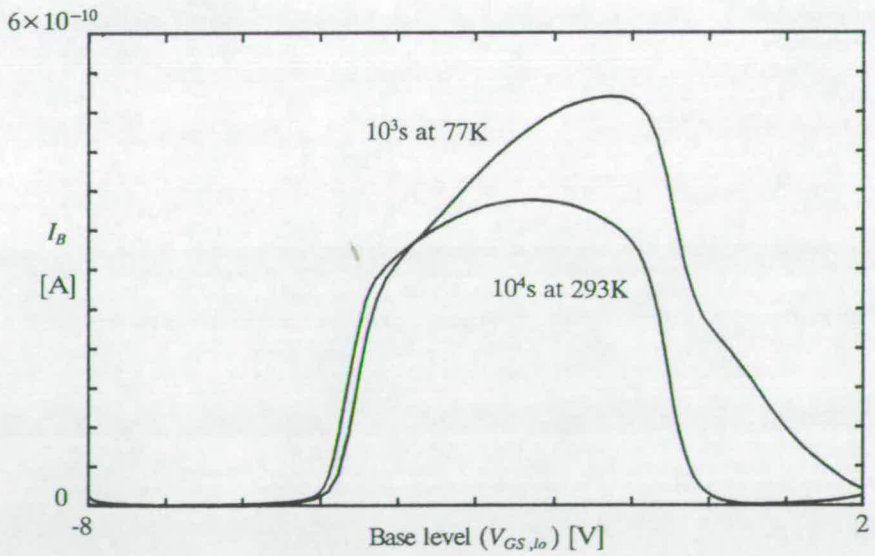


Figure 5.38: Comparison between charge pumping results of two 0.15 μm gap transistors. The first was stressed for 10000 seconds at room temperature. The second was stressed for 1000 seconds at 77 K.

same test/degradation/test routine used so far was then performed with the difference that the degradation step was run at 77 K instead of room temperature. The result obtained on a transistor with a 0.15 μm gap is depicted in Figure 5.38 before and after degradation, together with a charge pumping result obtained from another 0.15 μm-gap transistor degraded at room temperature. The drastic increase in $\overline{D_{it}}$ at low temperatures thus confirms the presence of hot carrier effects.

The test/degradation/test routine was then adapted to include the evaluation of the energy distribution of the interface states. The following procedure was used:

1) Transistor gap assessment using subthreshold current measurement ($V_{DS} = 1\ V$, $V_{GS} = 0$ to 5 V) in both direct and reverse configurations;

2) Charge pumping current measurement with source, drain and substrate grounded and gate pulses oscillating with an amplitude of 5 V at 250 kHz and a varying offset so that the base level goes from -8 V to 2 V;

3) Gated diode current measurement with source floating, $V_{DS} = 1\ V$, V_{GS} varying from -0.5 to 2.5 V and substrate grounded;

4) 1000 second degradation cycle with $V_{DS} = V_{GS} = 5\ V$ and source and substrate grounded;

5) Gated diode current measurement with source floating, $V_{DS} = 1\ V$, V_{GS} varying from -0.5 to 2.5 V and substrate grounded;

6) Charge pumping current measurement with source, drain and substrate grounded and gate pulses oscillating with an amplitude of 5 V at 250 kHz and a varying offset so that the base level goes from -8 V to 2 V;

7) Charge pumping measurement (for determining the energy distribution of interface states) consisting of applying a gate pulse as described in Figure 5.25 with a dead band T_d varying from 1 ns to 1 ms and grounding source and drain;

8) 10000 second degradation cycle with $V_{DS} = V_{GS} = 5\ V$ and source and substrate grounded;

9) Gated diode current measurement with source floating, $V_{DS} = 1\ V$, V_{GS} varying from -0.5 to 2.5 V and substrate grounded;

10) Charge pumping measurement (for determining the energy distribution of interface states) consisting of applying a gate pulse as described in Figure 5.25 with a dead band T_d varying from 1 ns to 1 ms and grounding source and drain;

11) Charge pumping current measurement with source, drain and substrate grounded and gate pulses oscillating with an amplitude of 5 V at 250 kHz and a varying offset so that the base level goes from -8 V to 2 V.

The conventional charge pumping and gated diode tests were retained in the procedure in order to supply more observations for the trend evaluation displayed in Figures 5.35 and 5.36. Like the first procedure, this was also repeated for several different groups of devices. Figure 5.39 shows a typical energy distribution result. A small increase in the peaks corresponding to the lowest and highest energies were observed in some cases. The most pronounced feature, however, was a decrease in the minimum densities at both lower and upper curves (below and above midgap).

The same test/degradation/test procedure was then repeated on a few more groups of transistors with a different bias condition for both the 1000 and 10000 second degradation cycles. V_{GS} was set to 1 V and V_{DS} to 5 V. As discussed in Chapter 3 this causes a predominance of hot hole injection. No significant change was observed in the charge pumping curves, except for transistors with 0.05 μm gaps or non-gapped transistors (Figure 5.40). No shifts towards more negative voltages were observed, thus indicating that there was no hole trapping (or that hole trapping was taking place concomitantly with electron trapping). On the contrary, the flatband voltage had a positive shift, similar to the hot electron injection case (Figure 5.41).

Finally, the test/degradation/test procedure was repeated on different groups with an a.c. gate bias during the stress cycles. A pulse with a base voltage of 0 V and a high level voltage of 5 V was applied to the gate with a 100 kHz frequency. Under a.c. stress conditions, when the gate voltage is at its maximum, hot electron injection prevails since $V_{GS} = V_{DS}$. When the transition from 5 V to zero occurs, there is a transient hole current to restore charge equilibrium [29,30] and those holes are heated by the high electric field pointing towards the gate. Additional degradation can also have been introduced by pulse overshoot caused by the test set-up.[31] Only hot electron injection was observed for large gaps up to a minimum gap limit of 0.15 μm (Figure 5.42.a). Below

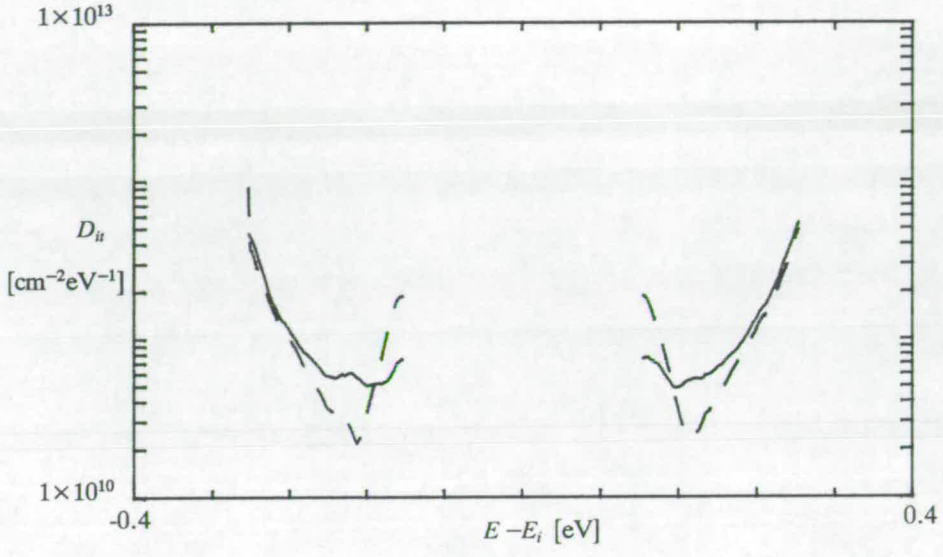


Figure 5.39: Energy distribution of interface states in a 1.5 μm transistor with a 0.20 μm gap before (solid line) and after (dashed line) degradation ($V_{GS} = V_{DS} = 5$V for 10000 seconds).

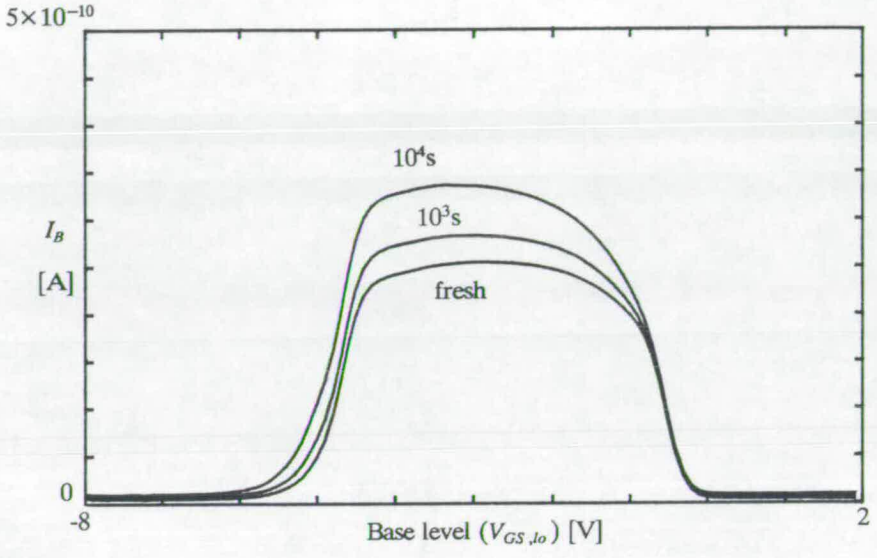


Figure 5.40: Charge pumping results for a 1.5 μm transistor with a 0.05 μm gap before and after 1000 and 10000 second stress routines ($V_{GS}=1$V, $V_{DS}=5$V).

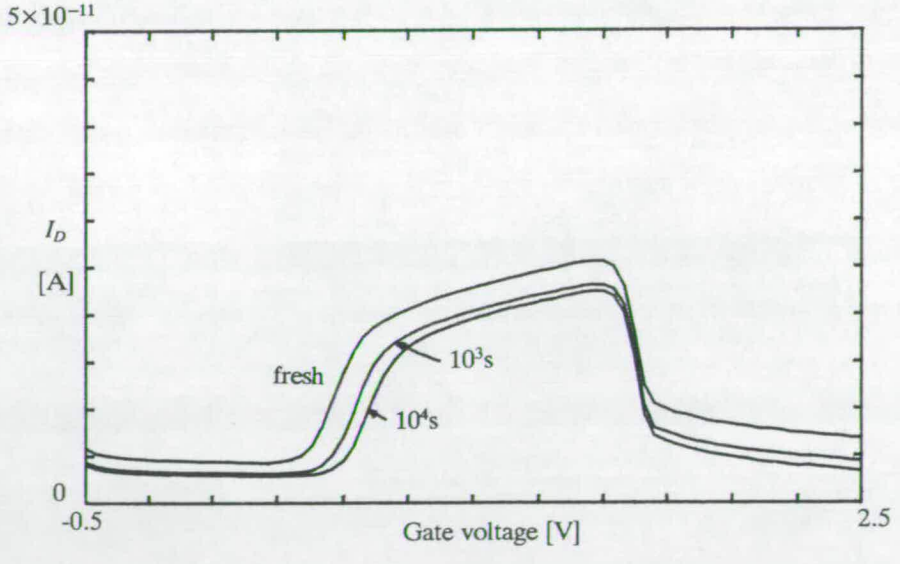


Figure 5.41: Gated diode results from the same transistor of Figure 5.40.

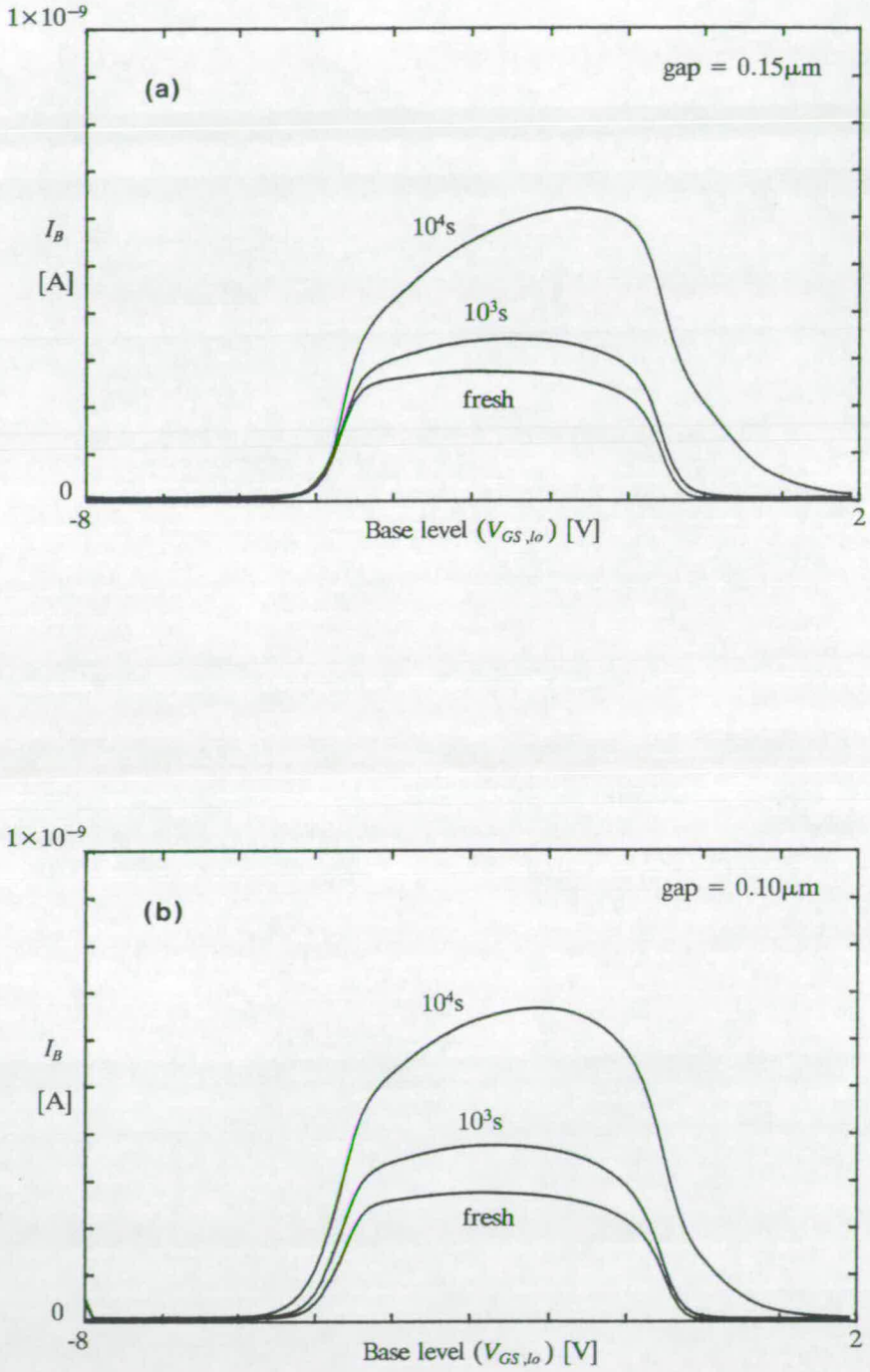
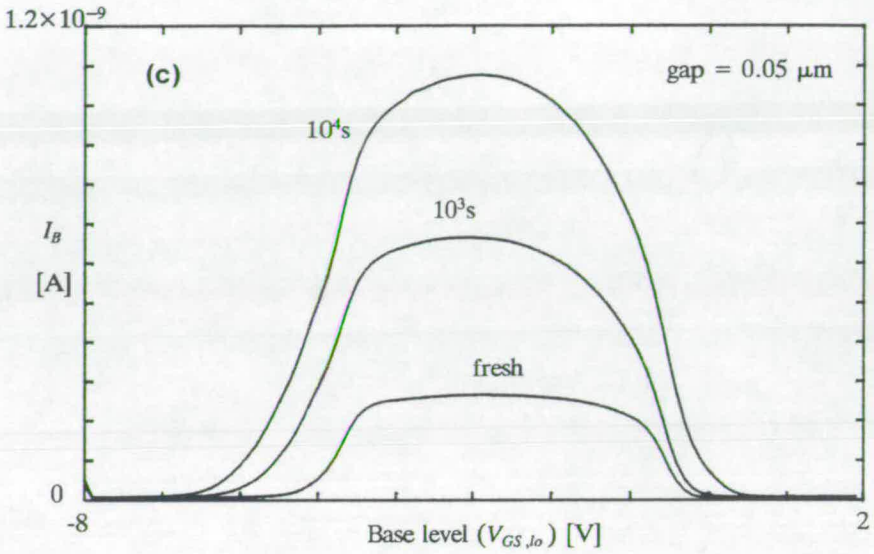


Figure 5.42: Charge pumping results of 0.15 μm transistors before and after 1000 and 10000 second stress routines (V_{DS} =5V, V_{GS} constituted by a square waveform oscillating between 0 and 5V at 100kHz).



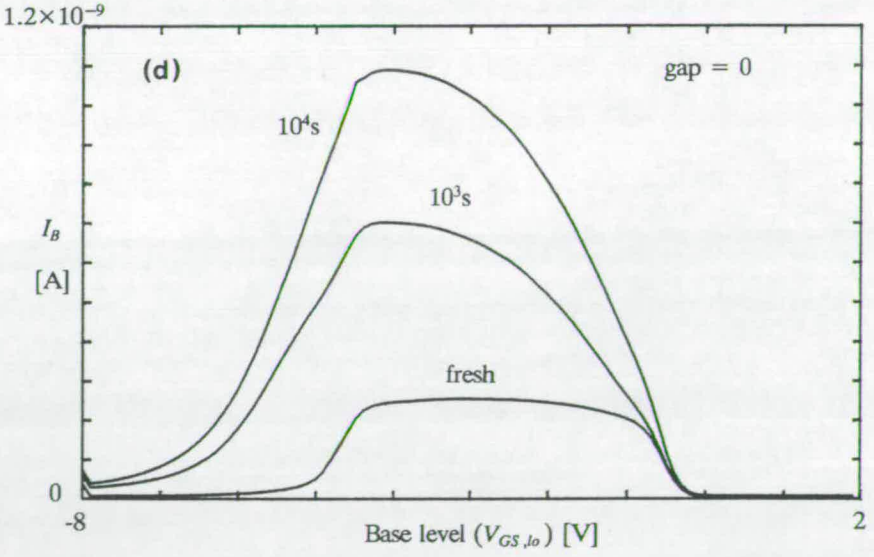


Figure 5.42 (cont'd)

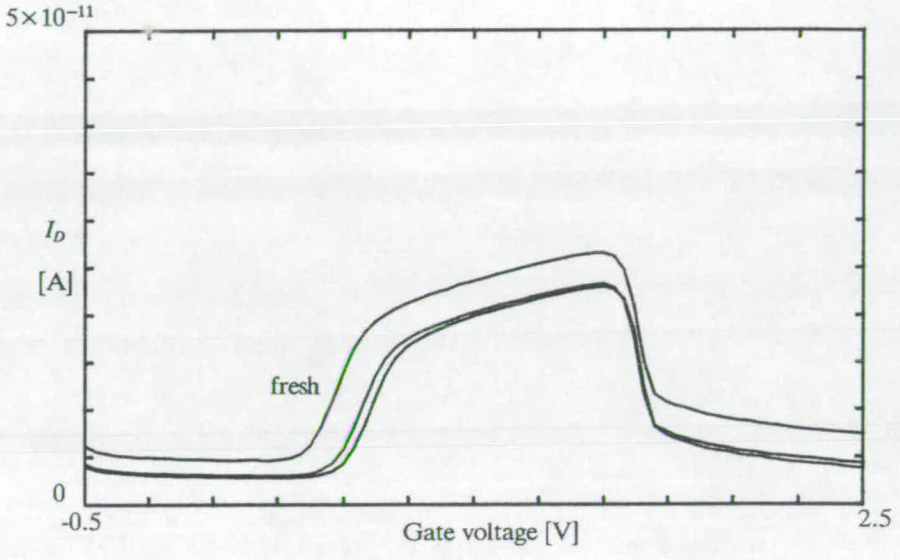


Figure 5.43: Gated diode results from the same zero-gap transistor of Figure 5.42.

that value, both hot hole and hot electron injection were observed, with the hot electron contribution decreasing as the gaps were decreasing (Figure 5.42.b and 5.42.c). When the gap was reduced to zero, only hot hole injection was detected (Figure 5.43.2). A gated diode measurement of this zero gap transistor (Figure 5.43) confirmed the absence of negative shift observed previously for the d.c. hot hole injection case.

References

1. Terada, K. and Muta, H., "A new method to determine effective MOSFET channel length," *Japan. J. Appl. Phys.*, vol. 18, pp. 953-959, May 1979.
2. Peng, K. L. and Afromowitz, M. A., "An improved method to determine MOSFET channel length," *IEEE Electron Device Lett.*, vol. EDL-3, pp. 360-362, December 1982.
3. Dennard, R. H., Gaensslen, F. H., Yu, H. N., Rideout, V. L., Bassous, E., and LeBlanc, A. R., "Design of ion implanted MOSFETs with very small physical dimensions," *IEEE J. Solid-State Circuits*, vol. SC-9, pp. 256-267, October 1974.
4. Pan, E. and Fang, F. F., "Lateral spread of ion-implanted impurities in silicon," *J. Appl. Phys.*, vol. 45, pp. 2801-2803, June 1974.
5. Wang, P. P., "Device characteristics of short-channel and narrow-width MOSFETs," *IEEE Trans. Electron Devices*, vol. ED-25, pp. 779-786, July 1978.
6. Chern, J. G. J., Chang, P., Motta, R. F., and Godinho, N., "A new method to determine MOSFET channel length," *IEEE Electron Device Lett.*, vol. EDL-1, pp. 170-173, September 1980.
7. Hu, G. J., Chang, C., and Chia, Y. T., "Gate-voltage-dependent effective channel length and series resistance of LDD MOSFETs," *IEEE Trans. Electron Devices*, vol. ED-34, pp. 2469-2475, December 1987.
8. Grove, A. S. and Fitzgerald, D. J., "Surface effects on p-n junctions: characteristics of surface space-charge regions under non-equilibrium conditions," *Solid-State Electron.*, vol. 9, pp. 783-806, 1966.

9. Sheu, B. J. and Ko, P. K., "A capacitance method to determine channel lengths for conventional and LDD MOSFETs," *IEEE Electron Device Lett.*, vol. EDL-5, pp. 491-493, November 1984.

10. Elliot, A. B. M., "The use of charge pumping currents to measure surface state densities in MOS transistors," *Solid-State Electron.*, vol. 19, pp. 241-247, 1976.

11. Roulston, D. J., *Bipolar Semiconductor Devices,* McGraw-Hill, New York, 1990.

12. Serack, J. A., Walton, A. J., and Robertson, J. M., "The effect of device geometry on IGFET characteristics," *ESSDERC Proc.*, pp. 915-918, Bologna, September 1987.

13. Hui, J., Hsu, F. C., and Moll, J., "A new substrate and gate current phenomenon in short-channel LDD and minimum overlap devices," *IEEE Electron Device Lett.*, vol. EDL-6, pp. 135-138, March 1985.

14. Chan, T. Y., Wu, A. T., Ko, P. K., Hu, C., and Razouk, R. R., "Asymmetrical characteristics in LDD and minimum-overlap MOSFETs," *IEEE Electron Device Lett.*, vol. EDL-7, pp. 16-19, January 1986.

15. Brugler, J. S. and Jespers, P. G. A., "Charge pumping in MOS devices," *IEEE Trans. Electron Devices*, vol. ED-16, pp. 297-302, March 1969.

16. Heremans, P., Witters, J., Groeseneken, G., and Maes, H. E., "Analysis of the charge pumping technique and its application for the evaluation of MOSFET degradation," *IEEE Trans. Electron Devices*, vol. ED-36, pp. 1318-1335, July 1989.

17. Groeseneken, G., Maes, H. E., Beltran, N., and De Keersmaecker, R. F., "A reliable approach to charge-pumping measurements in MOS transistors," *IEEE Trans. Electron Devices*, vol. ED-31, pp. 42-53, January 1984.

18. Hofmann, F. and Krautschneider, W. H., "A simple technique for determining the interface-trap distribution of submicron metal-oxide-semiconductor transistors by the charge pumping method," *J. Appl. Phys.*, vol. 65, pp. 1358-1360, February 1989.

19. Tseng, W. L., "A new charge pumping method of measuring Si-SiO_2 interface states," *J. Appl. Phys.*, vol. 62, pp. 591-599, July 1987.

20. Cilingiroglu, U., "Charge-pumping spectroscopy with pulsed interface probing," *IEEE Trans. Electron Devices*, vol. ED-37, pp. 267-272, January 1990.

21. Russell, T., Wilson, C. L., and Gaitan, M., "Determination of the spatial variation of interface trapped charge using short-channel MOSFETs," *IEEE Trans. Electron Devices*, vol. ED-30, pp. 1662-1671, December 1983.

22. Preier, H., "Contributions of surface states to MOS impedance," *Appl. Phys. Lett.*, vol. 10, pp. 361-363, 1967.

23. Cilingiroglu, U., "A pulsed interface-probing technique for MOS interface characterization at mid-gap levels," *IEEE Trans. Electron Devices*, vol. ED-35, pp. 2391-2396, December 1988.

24. Hu, C., Tam, S. C., Hsu, F., Ko, P., Chan, T., and Terrill, K. W., "Hot-electron-induced MOSFET degradation - model, monitor, and improvement," *IEEE J. Solid-State Circuits*, vol. SC-20, pp. 295-305, February 1985.

25. Hsu, F. C. and Grinolds, H. R., "Structure-enhanced MOSFET degradation due to hot-electron injection," *IEEE Electron Device Lett.*, vol. EDL-5, pp. 71-74, March 1984.

26. Pimbley, J. M. and Gildenblat, G., "Comments on "Structure-enhanced MOSFET degradation due to hot-electron injection"," *IEEE Electron Device Lett.*, vol. EDL-5, pp. 256-258, July 1984.

27. Muller, R. S. and Kamins, T. I., *Device electronics for integrated circuits*, Wiley, New York, 1977.

28. Kahng, D., "Silicon Integrated Circuits - Part A," *Applied Solid State Science (Supplement 2)*, Academic Press, New York, 1981.

29. Chen, K. L., Saller, S., and Shah, R., "The case of AC stress in the hot-carrier effect," *IEEE Trans. Electron Devices*, vol. ED-33, pp. 424-426, March 1986.

30. Fair, R. B. and Sun, R. C., "Threshold-voltage instability in MOSFETs due to channel hot-hole emission," *IEEE Trans. Electron Devices*, vol. ED-28, pp. 83-94, January 1981.

31. Bellens, R., Heremans, P., Groeseneken, G., Maes, H. E., and Weber, W., "The influence of the measurement setup on enhanced AC hot carrier degradation of MOSFETs," *IEEE Trans. Electron Devices*, vol. 37, pp. 310-313, January 1990.

Chapter 6 - Analysis of Results

6.1. Introduction

In this Chapter the significance of the introduction of gate offsets in terms of the alteration in the electric field distribution and in the density of available carriers is discussed. As seen in Chapter 5, the presence of such gaps made possible the detection of hot electron degradation in n-channel MOSFETs through the charge pumping technique. By analysing the modifications introduced by the gate offsets in the field and carrier density distributions and also by comparing results obtained from different techniques, it will be shown that it is also possible to determine the role of both types of carriers in degradation under the different bias conditions used in the experiments.

6.2. Significance of the offset gate experiment

In the preceding Chapter, it was seen that the introduction of a gate offset dramatically alters the hot carrier characteristics of a transistor. Hot electron degradation in particular becomes noticeable, in contrast with the situation for non-gapped devices. In Chapters 2 and 3 the conditions which lead to hot carrier generation and injection were reviewed, and from that it can be concluded that the two most significant factors are the density of available carriers and the distribution and magnitude of the electric field. The shift in the gate position obviously alters the field configuration and has been observed to degrade transistor characteristics, especially the current driving capability.[1] It is therefore important to know how the hot carrier generation conditions are affected as the gap is varied in order to relate the results to non-gapped transistors. We shall discuss in Sections 6.2.1 through 6.2.3 the conditions for each of the stress conditions used throughout the experiments of Chapter 5.

6.2.1. Case 1: $V_{GS}=V_{DS}=5$ V (hot electron degradation)

As we have seen previously in Chapter 3, the generation of hot carriers under such biasing conditions is caused primarily by the peak in the horizontal component of the electric field near the drain junction, particularly in the case of small geometries. The horizontal electric field stays more or less constant at low values throughout most of the channel length. Together with the horizontal field, it can define a preferential direction for the hot electron injection, although this is not a necessary condition considering that the direction of the hot electron injection is also strongly influenced by collision processes.

The presence of the gate on top of the edge of the drain junction and biased at the same voltage as the drain has an effect of spreading the potential lines. As a gap is formed between gate and drain, this effect gets increasingly weaker. Figure 6.1.a and 6.1.b depict the simulated potential distribution of 1.5 μm transistors with zero and 0.35 μm gaps respectively; a more pronounced crowding of potential lines can be noticed in the latter case. The consequence of this potential crowding is an increase in the electric field as the gap size increases, as demonstrated by Figures 6.2.a and 6.2.b which show the variation of the horizontal and vertical components of the electric field at the surface, corresponding to Figure 6.1.a and 6.1.b, respectively. The variation of the peak of the electric field with the gap size is given in Figure 6.3. As expected, it increases with the gap size and tends to saturate as the gate influence upon the field near the drain becomes weaker. The increase in the peak of the electric field is not accompanied by a comparable increase in the electric field gradient; therefore the field values near the peak are also increased. This means that there is an increase in the length of the region near the drain where injection takes place. *This is how hot electron injection becomes detectable by the charge pumping technique.*

As the gate is moved away from the drain, it can no longer invert the whole channel. A depletion layer is then formed in the gap region and the conduction which was previously confined to the narrow inverted channel near the interface spreads over a much wider region. This situation is similar to the case analysed in Chapter 3 when the transistor is biased into saturation, with the important difference that in this case the vertical field is not favourable to hot hole injection. Indeed, for a non-gapped transistor,

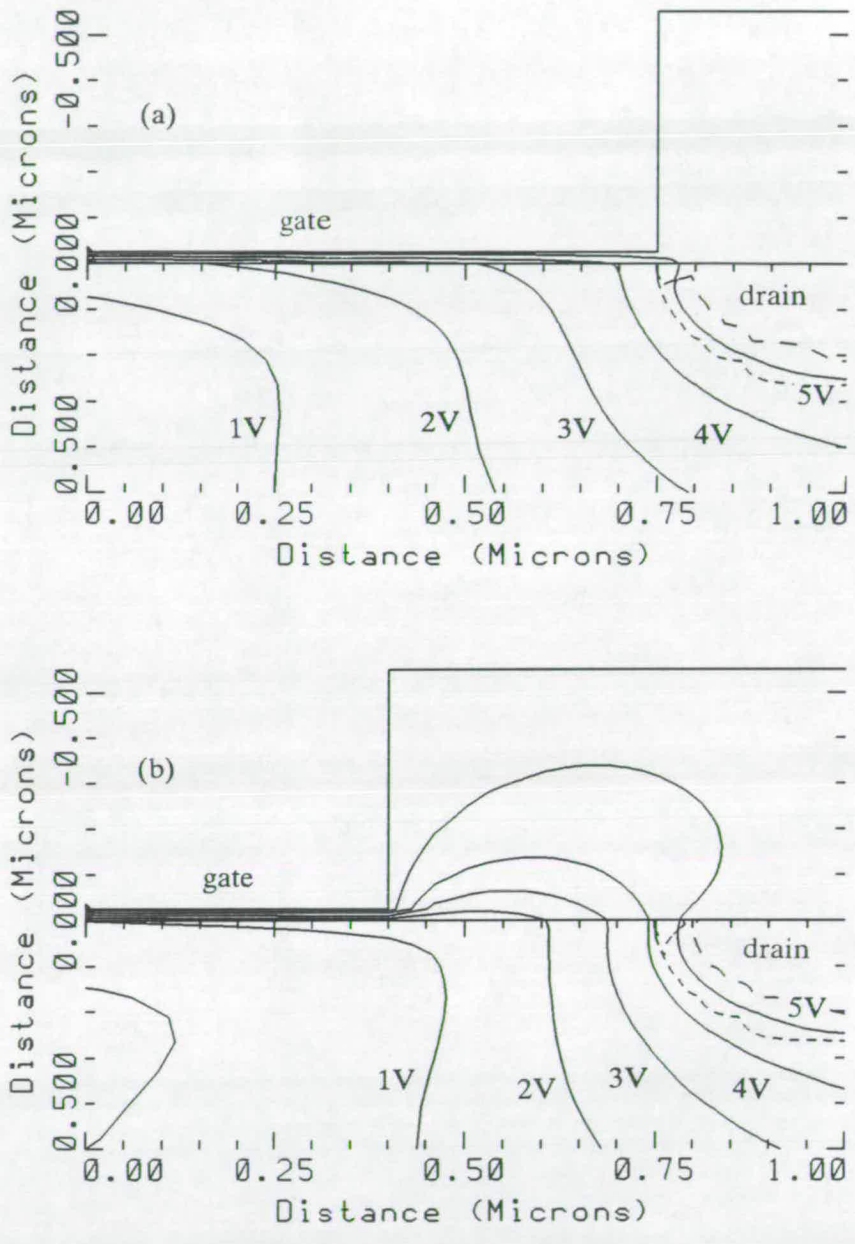


Figure 6.1: Simulation of the potential distribution (solid lines) near the drain junction of 1.5 μm transistors with (a) zero gap and (b) 0.35μm gap. $V_{GS} = V_{DS} = 5V$. The increment between lines is 1V.

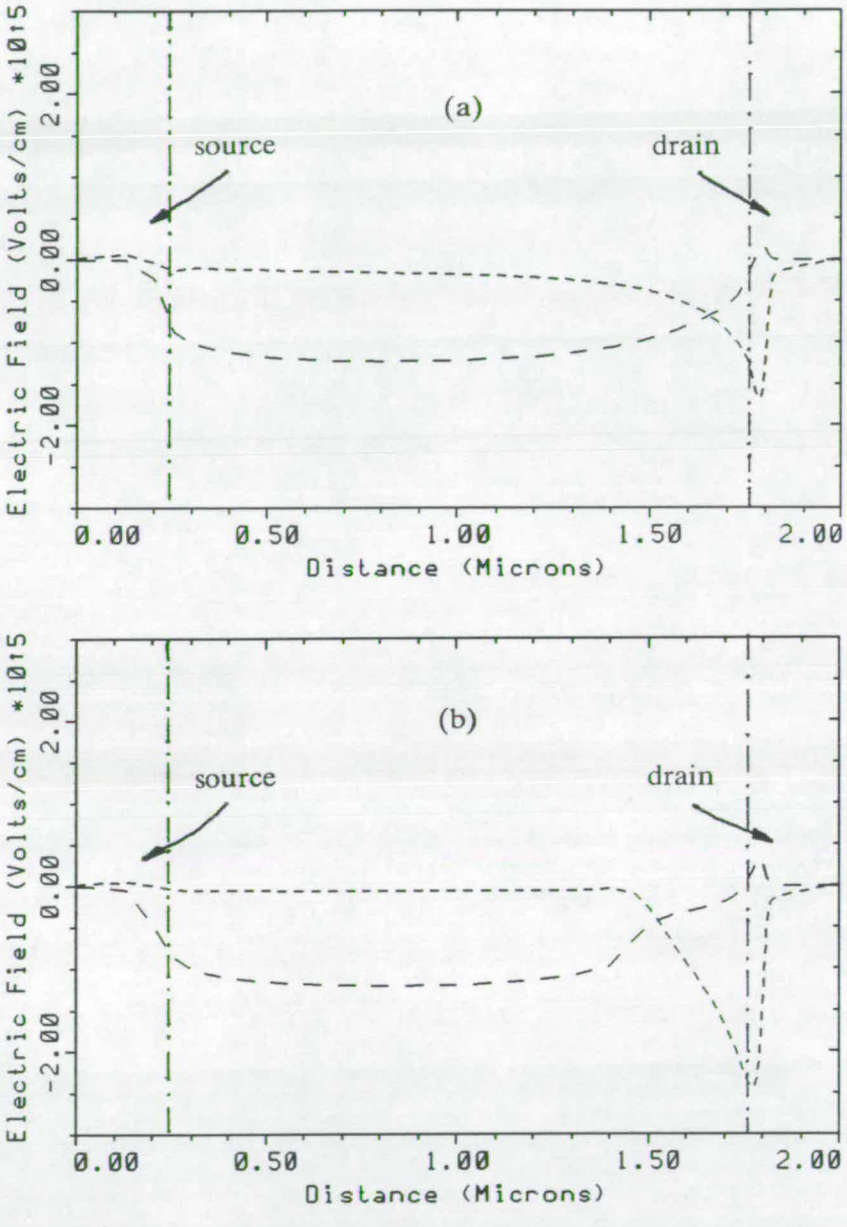


Figure 6.2: Horizontal (short-dashed line) and vertical (long-dashed line) components of the electric field across the channel. (a) zero gap. (b) 0.35 μm gap. $V_{GS} = V_{DS} = 5V$.

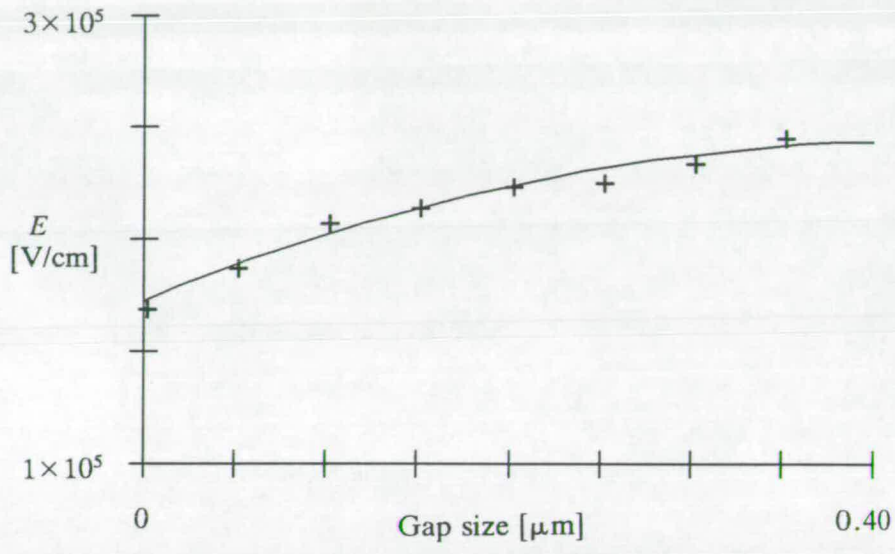


Figure 6.3: Simulation of the variation of the horizontal field peak versus the gap size (1.5 μm channel length, $V_{GS} = V_{DS} = 5V$).

the reduction of the gate voltage (with a fixed drain voltage) has the effect of reducing the length of the inverted region; the other way to do it is to move the gate away from the drain. The occurrence of pinch-off under this "geometry-induced saturation" is confirmed in Figure 6.4.

The reduction of what could be regarded as the "effective gate voltage" also has an effect on the density of available carriers. With the increase in the gap size, the drain saturation current becomes increasingly low. Figure 6.5 shows the measured drain current for a group of transistors with gaps ranging from 0 to 0.35 μm. It shows that, from the point of view of the drift current, the availability of carriers for the heating process becomes increasingly small. However, it should not be forgotten that as the depletion layer is formed and expanded as a consequence of the saturation process, impact ionisation also starts to contribute more carriers to be eventually heated. Therefore the fall in the density of generated hot carriers may not be as sharp as the decrease in saturation current.

It can therefore be concluded that an increase in hot electron degradation should be expected as the gap is increased, due to the variation in the electric field. However, an accurate analysis of the rising edge of Figure 5.34 is difficult because there are at least three different effects involved: 1) the increase, although small, in the electric field in the region near the junction; 2) the variation in the electric field gradient as the gap is increased, which affects the electron temperature; and 3) the detectability of the charge pumping current as a function of the distance between the participating interface states and the drain junction. As the gaps become larger, there can be a decrease in hot electron generation (and consequently injection) because of the reduction in the density of carriers available for heating. Looking back at the curve of Figure 5.34 (variation in the density of interface states versus gap size), this could perhaps explain the fall in the density of interface states for larger gaps, despite the fact that the real magnitude and the onset of the fall in the density of carriers available for heating is difficult to estimate because of the contribution of the impact ionisation component. However, as can be seen in Figure 5.34, the fall in ΔD_{it} does not correspond to a fall in trapped charge. Therefore the fall in the generation of new interface states cannot be explained by a decrease in carrier density and is probably related to the mechanisms involved, This is

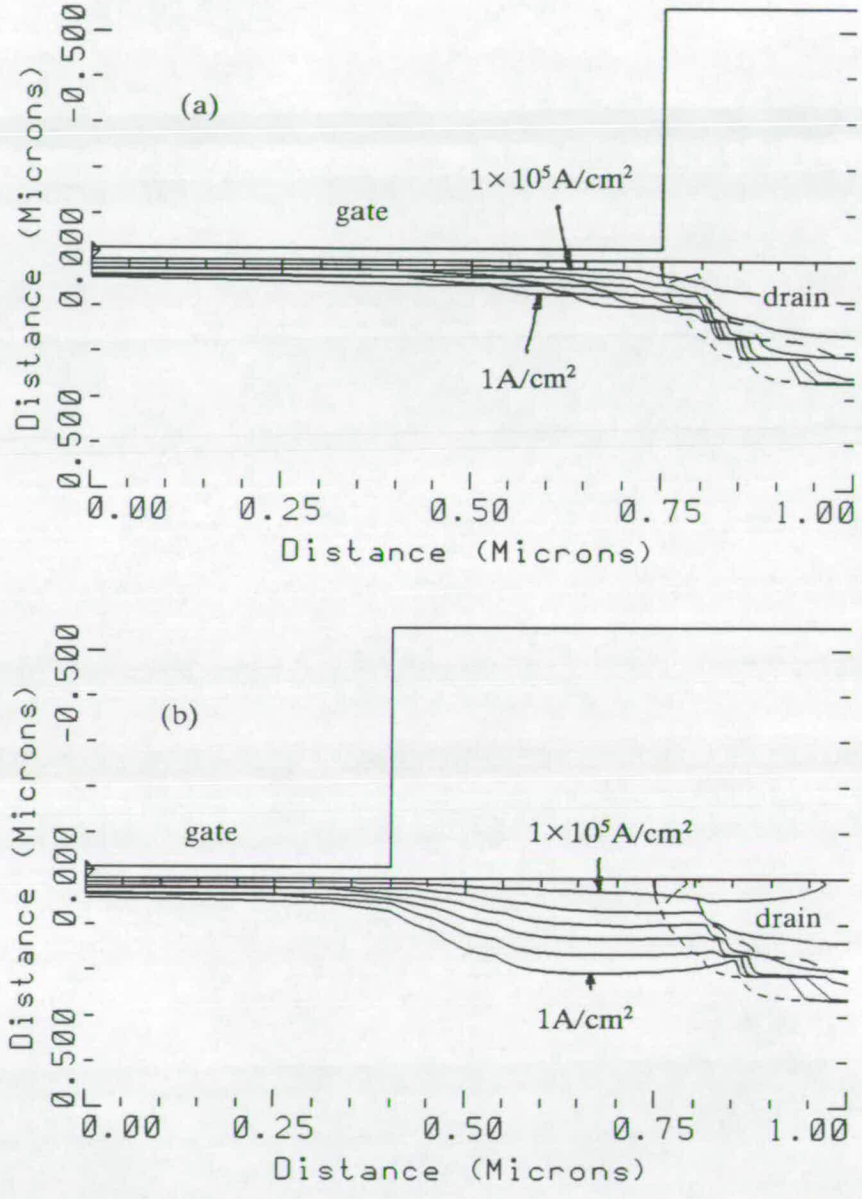


Figure 6.4: Simulated electron current density (solid lines) close to the surface near the drain diffusion of a 1.5 μm transistor. (a) zero gap. (b) 0.35 μm gap. $V_{GS} = V_{DS} = 5V$. The ratio between lines is 10.

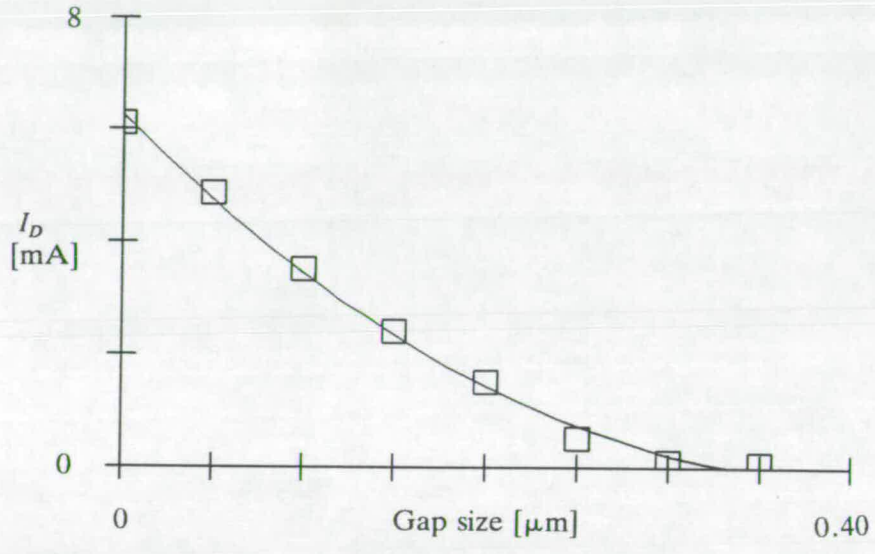


Figure 6.5: Measured drain current versus gap size of 1.5 μm transistors. $V_{GS} = V_{DS} = 5V$.

discussed later in Section 6.4.

6.2.2. Case 2: $V_{GS}=1\,V$, $V_{DS}=5\,V$

As seen in Chapter 3, the main conditions for the predominance of hot hole injection are the presence of a depletion region with a high electric field so that holes can be generated through impact ionisation and become heated and the presence of a vertical electric field pointing upwards. In the case of hot electron generation discussed in Section 6.2.1, the electrons were being supplied both directly by the drift current and by the impact ionisation process. In this case the only source of holes is impact ionisation. However, 5 V is too low a value to result in a high density of hot holes.

For a zero gap, the gate biased at a lower voltage than the drain has the effect of increasing the electric field near the drain edge. This can be verified by comparing Figures 6.6.a and 6.6.b, which show the simulation of the electric field distribution in 1.5 μm transistors with zero and 0.35 μm gaps, respectively. The reason for this increase can be understood if we look at the distribution of equipotential lines. For a gate situated far from the drain junction, the potential drops gradually with distance (Figure 6.7.b). When they are close together, the potential difference between them implies a voltage drop over a short distance, therefore leading to a crowding of equipotential lines between drain and gate (Figure 6.7.a). This crowding occurs both horizontally and vertically, thus causing a high horizontal and vertical field (Figure 6.8.a). As the gap is increased, the vertical electric field in the region near the drain junction falls abruptly to values near zero because of the large gap increments compared to the thickness of the gate insulator (Figure 6.8.b). In that case the vertical field no longer favours hole injection, although either holes or electrons can be directed towards the insulator through collisions. Therefore it leads to a situation similar to the $V_{GS}=V_{DS}=5V$ case for large gaps studied before, as we can see by comparing Figures 6.8.b and 6.1.b. The essential difference is in terms of the availability of carriers.

In this situation the fall in the density of carriers available for heating is much more than in the previous case, because the gate voltage is already low even at zero gap. Even though the field is still high for large gaps, the drift current is too low. No

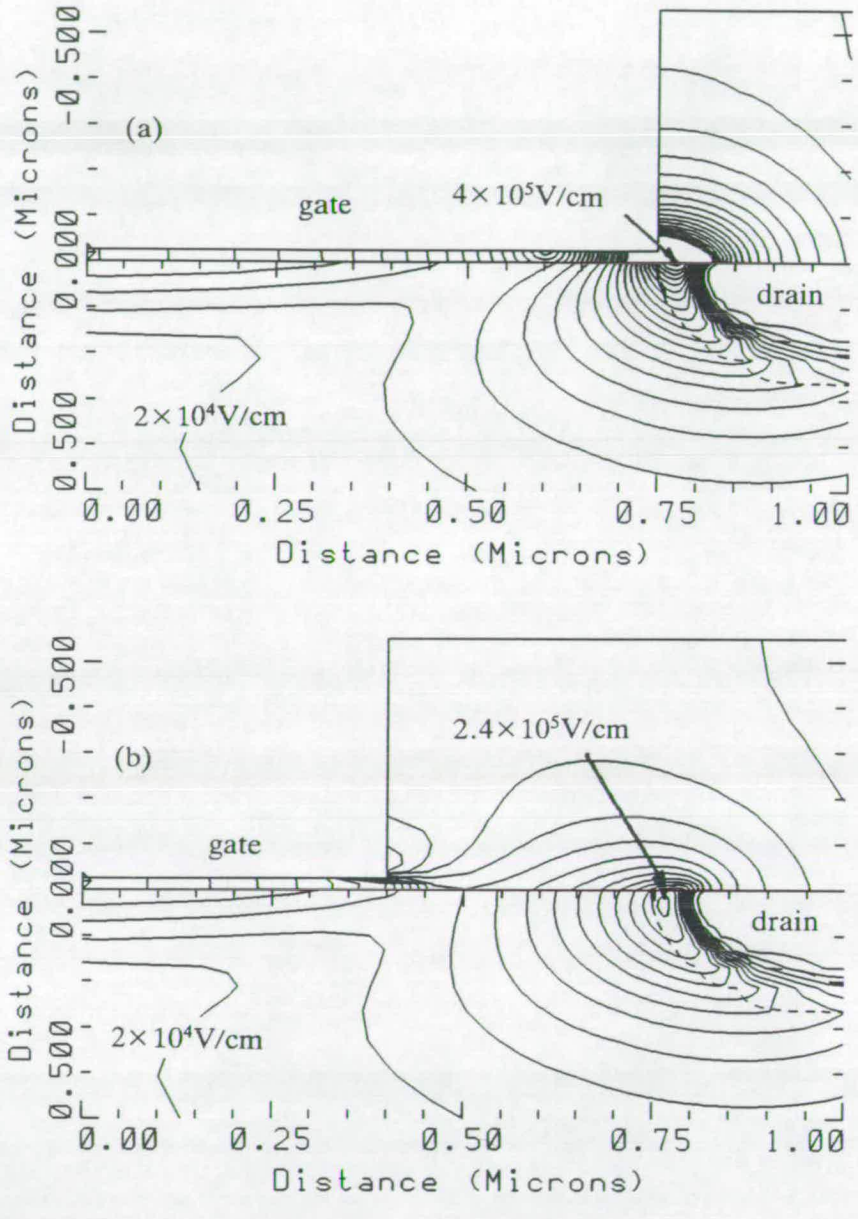


Figure 6.6: Simulation of the electric field distribution (solid lines) near the drain junction of 1.5 μm transistors with (a) zero gap and (b) 0.35μm gap. $V_{GS}=1$V, $V_{DS}=5$V. The increment between lines is 1×10^4V/cm.

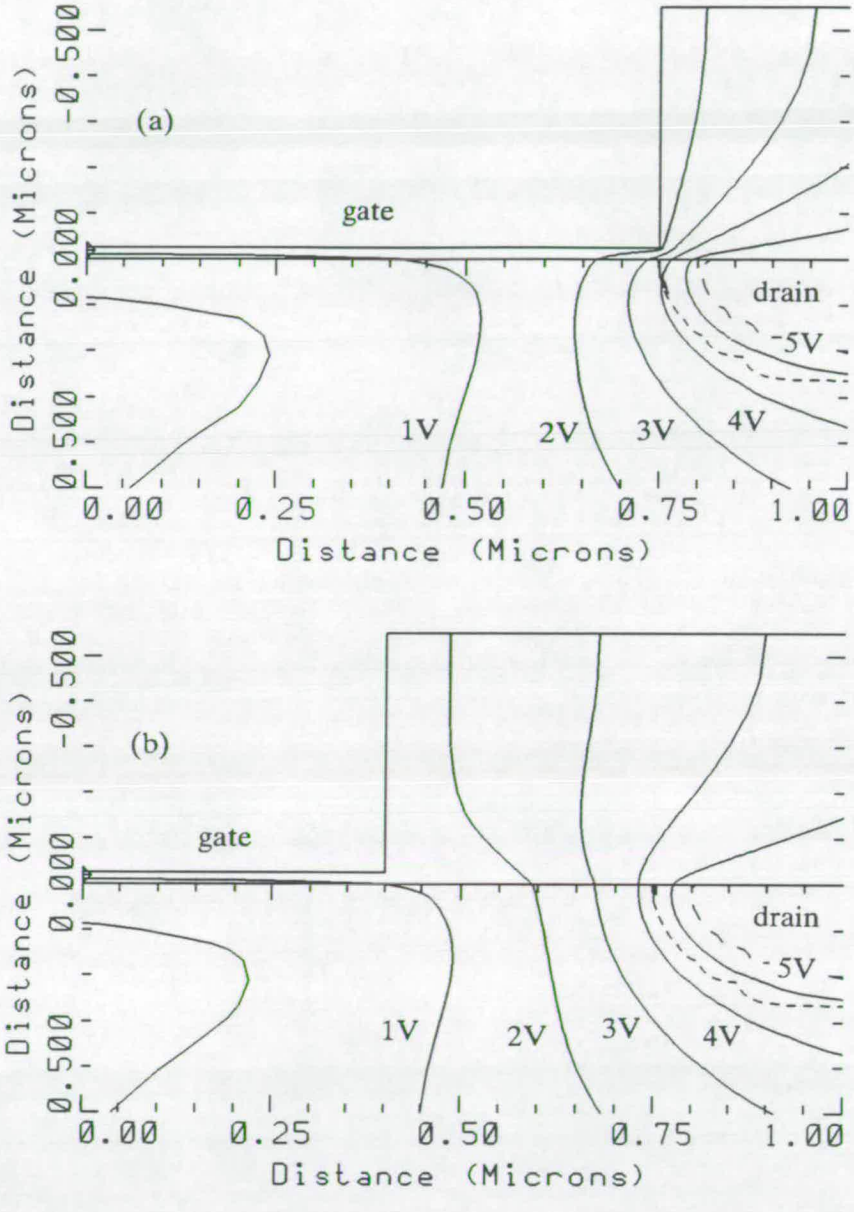


Figure 6.7: Simulation of the potential distribution (solid lines) near the drain junction of 1.5 μm transistors with (a) zero gap and (b) 0.35μm gap. $V_{GS}=1V$, $V_{DS}=5V$. The increment between lines is 1V.

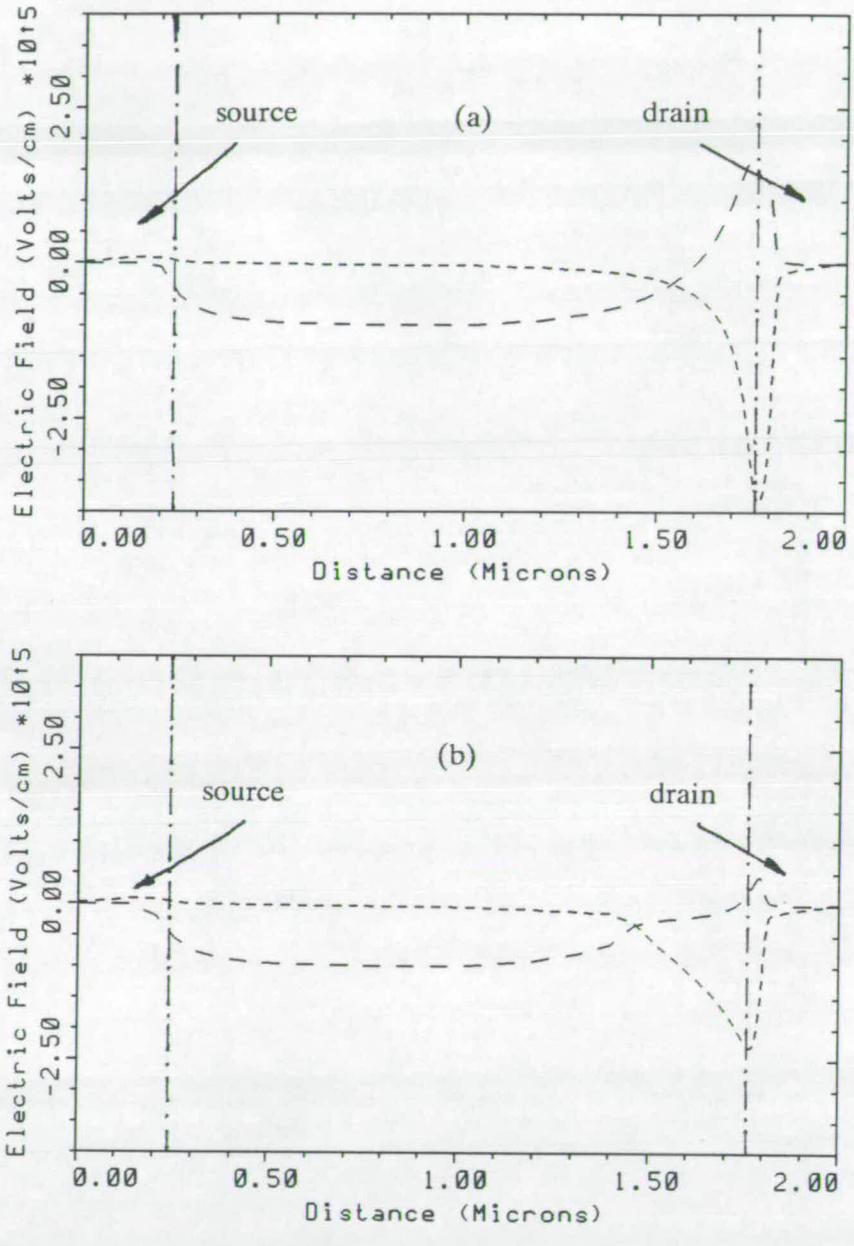


Figure 6.8: Horizontal (short-dashed line) and vertical (long-dashed line) components of the electric field across the channel. (a) zero gap. (b) 0.35 μm gap. $V_{GS} = 1V$, $V_{DS} = 5V$

appreciable degradation effect of either type is then observed.

6.2.3. Case 3: V_{DS}=5 V, V_{GS}= 0 to 5 V at 100 kHz (a.c. degradation)

In this case, both hot electrons and hot holes can be generated and injected. The process of hot electron generation is identical to the d.c. case. The generation of hot holes, however, is not initiated by impact ionisation. As discussed briefly in Chapter 5, holes come from the bulk towards the surface in order to restore charge neutrality when the gate pulse goes from 5 V to zero and can also be generated by pulse overshoot. They depend on the gradient of the vertical electric field to be heated. A comparison between the electric field distributions for a zero gap and a 0.35 μm gap transistor (Figures 6.6.a and 6.6.b) shows that the high gradient present in the zero gap situation is turned into purely vertical field lines in the large gap case. On the other hand, the generation of hot electrons depends on the horizontal field, in the same way as discussed for the $V_{GS} = V_{DS} = 5V$ case. The predominant degradation is therefore due to hot electrons.

As the gap is decreased, hot hole generation becomes possible and both types of carriers contribute to degradation. When the gap size approaches zero, the charge pumping technique is no longer capable of detecting hot electron degradation because of position of the injection region (too close to the drain junction) and only the hot hole degradation component can be measured.

6.3. Predominance of hot electron versus hot hole degradation

From what has been discussed above and considering the results presented in Chapter 5, we can conclude that there is a prevalence of hot electron degradation in gapped transistors both under the linear and the saturation regime. In the latter case, it happens only when the vertical field does not favour hot hole injection. From the field analysis and based on the results obtained from the various offsets, the only situation when the contribution of hot electron degradation was not detected was under the d.c. saturation regime. Even in that case, a slight positive shift was repeatedly detected in the gated diode measurements, thus indicating the presence of a net negative charge.

The presence of negative trapped charge after hot hole injection in IGFETs was explained by Hofmann[2] using a model developed by Lai[3] from experiments with MOS capacitors. According to this model, interface traps are generated when holes are injected. If this process is followed by an electron injection, the electrons can be trapped by the holes themselves and generate new interface states. The mechanisms involved are still unclear, but Lai suggested that electrons and holes recombine and release energy which in turn generates new interface states. Such interface states are acceptor-type states. We must observe, however, that if this were the case, the charge pumping curve would show the presence of a positively shifted component introduced by such acceptor-type states as well as the negatively shifted component caused by the donor-type interface states generated by the holes. In the a.c. degradation case, for transistors with gaps between 0.15 μm and zero, the presence of both acceptor and donor-type interface states could be detected by the charge pumping method, thus indicating that the method would be able to detect the generation of acceptor-type interface states by the electron-hole recombination process. Therefore there must be a different explanation for the negative charge trapped under conditions of hot hole injection.

No such discrepancy between the type of interface state and the type of charge was observed in the hot electron injection case. There was a positive shift in the gated diode curve, thus indicating the presence of negative trapped charges and the charge pumping curve showed a degradation component shifted positively, which corresponds to the presence of acceptor-type interface states. Quantitatively, however, it was observed that the decrease in the average density of generated interface states for larger gaps (above 0.15 μm) was not followed by the decrease in trapped charge obtained from the gated diode measurements. This phenomenon can also be observed by comparing the charge pumping curves for different gap sizes. When the gap is increased from 0.15 to 0.20 μm, for example (Figure 6.9), there is a decrease in the peak of the charge pumping curve and an increase in the lateral spread of its falling edge. For larger gaps, such lateral spread cannot be easily distinguished because of the low amplitude of the degraded component of the charge pumping current.

The contrast between the increase in trapped charge and the decrease in the density of interface states for gaps larger than 0.15 μm, together with the observation of trapped

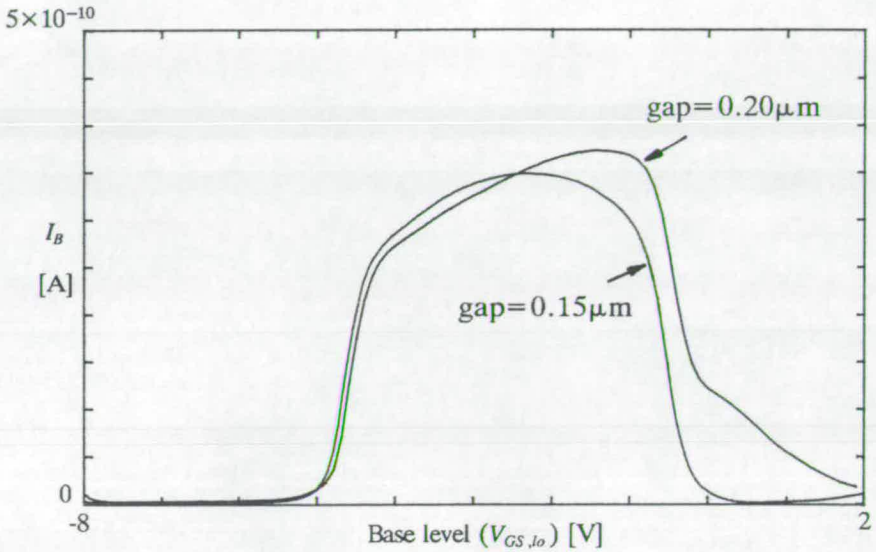


Figure 6.9: Comparison between the charge pumping curve of a 0.15μm-gap and a 0.20μm-gap transistor. They were both stressed with $V_{GS} = V_{DS} = 5\text{V}$ for 10000 seconds.

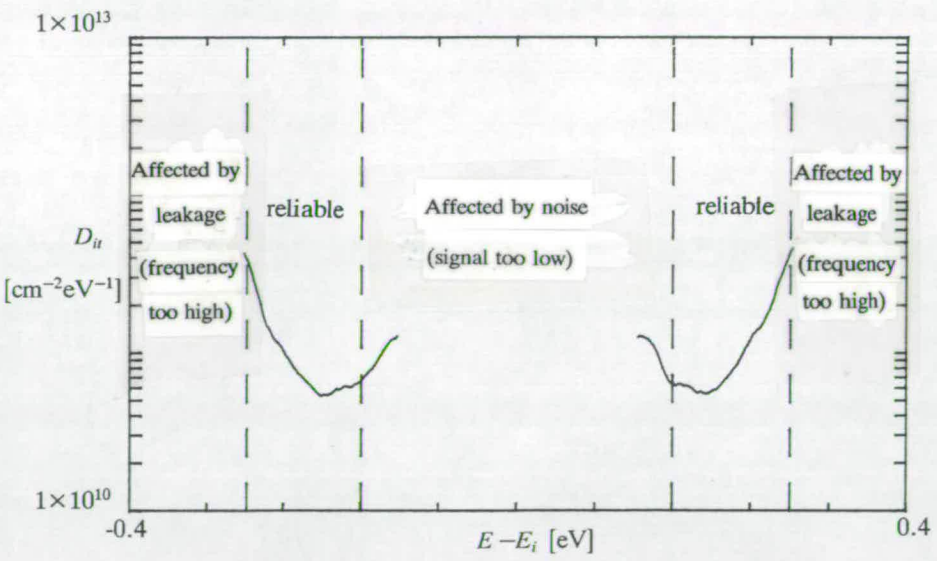


Figure 6.10: Example of measured energy distribution of interface states. The curve was divided into different zones according to the confidence of data.

negative charge not associated with the generation of interface states in the hot hole injection case seems to suggest a common cause. Other information can be added to this evidence: the reduction in trapped charge observed over long periods. As discussed in Chapter 5, samples were left resting for long periods of time (72 hours) after hot electron injection in order to confirm the detection of negative trapped charge by the charge pumping measurements. A negative shift was observed in the falling edge of the charge pumping curve, indicating the reduction in negative charge. All these results seem to indicate the trapping of negative charge in the bulk of the SiO_2, not related to the interface states. In the case of hot electron generation, the injected carriers would then be responsible for both the generation of acceptor-type interface states and of slow traps. Under hot hole injection conditions, the density of generated electrons decreases sharply, but they would still be generated by the same impact ionisation process which generates hot holes and be injected into the SiO_2, where they undergo the same acceptor generation and bulk trapping as in the previous case. However, these acceptor-type interface states would be by far outnumbered by the generation of donor-type interface states by injected holes. This is because hot holes are much more efficient in generating interface states than hot electrons.[4] Under a.c. stress, there is an increase in acceptor-type interface state generation, which then becomes comparable to the generation of donor-type interface states. This is caused by the greater availability of hot electrons, which are generated during the high level of the gate pulse. This would explain why an increase in both types of interface states was detected.

6.4. Energy distribution of generated interface states

The use of the adapted charge pumping method (described in Chapter 5) for the determination of the energy distribution of interface states inevitably leads to some limitations. The variation of the transition times of the gate signal either by varying the slopes (Groeseneken's method) or by introducing a dead band (Hofmann's method) has to be done over a very broad range due to the logarithmic dependence of the scanned energy on the transition times (Equation 5.2). Therefore the frequency of the resulting waveform applied to the gate has to vary broadly. As we have seen in Chapter 3, the charge pumping current is proportional to the frequency of the gate signal (Equation

3.7). If the frequency has to reach very low values, the current is low and the introduction of noise in the computation of the density of interface states is inevitable. This is the case for energy values situated near the midgap. Therefore the data points calculated in that region are not reliable. On the other hand, energies very close to the conduction and valence bands are difficult to scan because very short transition times become necessary. Short transitions imply in the presence of high frequency harmonics, which are prone to attenuation and leakage introduced by the system set-up. Therefore the increase of D_{it} observed for energies near midgap is likely to have been caused by measurement noise. For non-degraded devices, the density of interface states should remain more or less constant in this region; therefore the region near midgap was disregarded in the present analysis. The significance of the different regions of the D_{it} versus $E - E_i$ is summarised in Figure 6.10.

In the course of this work, charge pumping measurements were always accompanied by gated diode tests. It is therefore desirable to know whether the gated diode curve conveys any information on the energy distribution of interface states in order to confirm results obtained from the adapted charge pumping method. As discussed in Chapter 3, the current of a reverse biased gated diode can be used to assess the presence of interface states through carrier generation. This is essentially different from the mechanisms involved in the charge pumping phenomenon, where the process of filling and emptying interface states is forced by the pulse transitions. Because there is no control over the capture and emission of carriers, the energy distribution of the interface states which contribute to the interface component of the gated diode current (labeled I_{int} in Figure 3.15) is determined solely by the statistics of the generation process (in this case generation is the dominant process because of the low carrier concentration in the depletion region). This statistics are similar to the bulk generation/recombination statistics derived by Shockley, Hall and Read (usually referred to as SHR process).[5] According to this procedure, the recombination/generation rate is given by

$$U = \frac{D_{it} v_{th} \sigma (pn - n_i^2)}{p + n + 2n_i \cosh\left(\frac{E_{it} - E_i}{kT}\right)} \qquad (6.1)$$

where E_{it} is the energy of the interface state. As E_{it} approaches E_i, the denominator decreases and U increases. Therefore interface states closer to midgap will be more likely to contribute to the gated diode current. The interface states have to be empty for the generation process to occur. Under non-equilibrium conditions, centres will be empty of holes if the hole quasi-Fermi level is above the intrinsic level; similarly, they will be empty of electrons when the electron quasi-Fermi level is below the intrinsic level. As pointed out by Speckbacher,[6] a variation in gate potential will simply shift the spatial region for which the emptiness condition discussed above is valid. Therefore such variation in gate bias will not change the "energy window" of contributing interface states but will only be able to shift the "space window" across the channel at which surface generation is taking place.

As depicted in Figure 6.11, there was a decrease in the density of interface states for values in the reliable zone (as defined by Figure 6.10) of the D_{it} versus $E - E_i$ curve under the hot electron injection condition, despite the clear increase in the average density of interface states, detected by the peak of the conventional charge pumping curves. On the other hand, gated diode results showed a decrease in the interface generation current component, as shown in Figure 6.12 in the case of large gaps. Both results (modified charge pumping and gated diode curve) seem to indicate a decrease in the density of interface states near midgap. A possible explanation for this phenomenon is that the injected electrons modify the structure of the existing interface states, thus shifting them in the energy spectrum. Considering that the midgap is being monitored by the gated diode technique and the intermediate regions (between midgap and the conduction and valence bands) are being observed by the modified charge pumping method, only the energy levels near the conduction and valence bands were not monitored. Any increase in the density of interface states in those regions would remain unoticed. It therefore suggests that deep interface states would then be shifted to shallower energy levels, where neither the adapted charge pumping technique (because of the low frequency/low current limitations already discussed) nor the gated diode method (because of its near-midgap characteristic) would be able to detect them.

Finally, it is important to point out that the detection of midgap interface states by the gated diode method was only possible because of the presence of a drain gap. When

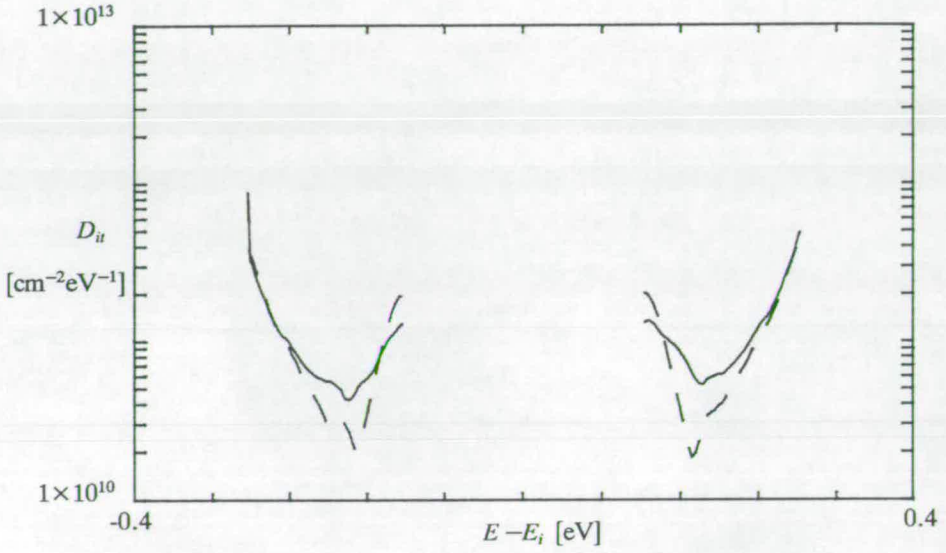


Figure 6.11: Energy distribution of interface states before (solid line) and after (dashed line) a 10000 second stress ($V_{GS} = V_{DS} = 5\text{V}$). The transistor has a 0.20μm drain gap.

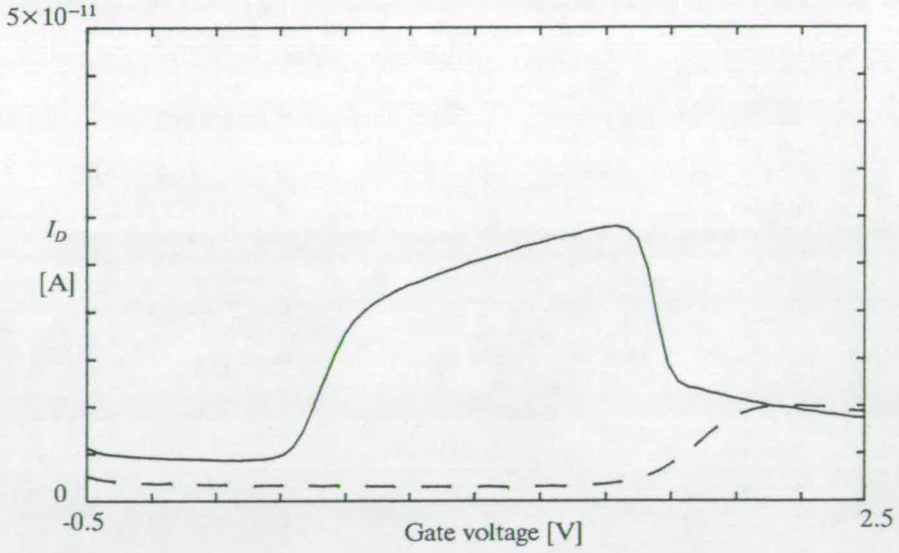


Figure 6.12: Gated diode results corresponding to Figure 6.11.

no gap is present, the gate voltages at which the spatial scanning of the interface states near the drain take place are below the channel flatband condition. This means that when the channel depletion region is formed, the region near the drain was already scanned by lower voltages; therefore those regions can only contribute to the I_{acc} current component (shown in Figure 3.15). However, when a gap is present, the effect of the gate voltage on the region near the drain is weakened by the gap. When the channel is depleted the region adjacent to the drain junction is still being scanned. This explains why the reduction in I_{int} was not detected in small gap transistors.

6.5. Summary of conclusions

The use of offset transistors in this work proved to be successful in detecting the presence of hot electron degradation in n-channel IGFETs through the charge pumping technique. This observation had not been possible before except when multi-layered gate insulators (including Si_3N_4 [7]) were used to boost the trapping process. In the present case, it has been made possible using a single layer of thermally grown SiO_2 by increasing the electric field in the vicinity of the junction, therefore making injection possible further from it, where it becomes detectable by the charge pumping method.

The results obtained from from transistors with different gaps submitted to different stressing conditions made it possible to isolate hot hole and hot electron effects. Furthermore, data gathered from charge pumping in its standard form, from the modified technique for the determination of the energy distribution of interface states and from the gated diode method have been brought together to confirm the nature of degradation in each case and to try to explain the mechanisms involved. However, the very differences in device operation and field distribution introduced by drain gaps made impossible the quantification of the degradation phenomena in order to establish a relationship between the effects in gapped and normal transistors. Nevertheless, the set of results made it possible to suggest qualitative models for charge trapping and the energy distribution of the generated interface states. In the first it is proposed that electrons are trapped in the bulk of SiO_2 either when they result from carriers participating of the channel drift current or generated through impact ionisation. In the latter case, the simultaneous presence of hot hole injection generated donor-type interface states but it does not affect the electron

bulk trapping process. The second model suggests that besides creating new interface states, hot electrons can also shift their energy, transforming deep level states into shallow states.

References

1. Chan, T. Y., Wu, A. T., Ko, P. K., Hu, C., and Razouk, R. R., "Asymmetrical characteristics in LDD and minimum-overlap MOSFETs," *IEEE Electron Device Lett.*, vol. EDL-7, pp. 16-19, January 1986.
2. Hofmann, K. R., Werner, C., Weber, W., and Dorda, G., "Hot-electron and hole-emission effects in short n-channel MOSFETs," *IEEE Trans. Electron Devices*, vol. ED-32, pp. 691-699, March 1985.
3. Lai, S. K., "Two-carrier nature of interface-state generation in hole trapping and radiation damage," *Appl. Phys. Lett.*, vol. 39(1), pp. 58-60, July 1981.
4. Heremans, P., Bellens, R., Groeseneken, G., and Maes, H. E., "Consistent model for the hot-carrier degradation in n-channel and p-channel MOSFETs," *IEEE Trans. Electron Devices*, vol. ED-35, pp. 2194-2209, December 1988.
5. Muller, R. S. and Kamins, T. I., *Device Electronics for Integrated Circuits,* Wiley, New York, 1986.
6. Speckbacher, P., Asenov, A., Bollu, M., Koch, F., and Weber, W., "Hot-carrier induced deep-level defects from gated-diode measurements on MOSFETs," *IEEE Electron Device Lett.*, vol. 11, pp. 95-97, February 1990.
7. Heremans, P., Witters, J., Groeseneken, G., and Maes, H. E., "Analysis of the charge pumping technique and its application for the evaluation of MOSFET degradation," *IEEE Trans. Electron Devices*, vol. 36, pp. 1318-1335, July 1989.

Chapter 7 - Conclusions

7.1. General remarks

The use of offset-gate transistors proved to be a valuable tool for the analysis of degradation effects in IGFETs. By increasing the electric field in the vicinity of the drain junction, the observation of hot electron injection using the charge pumping method became possible. The role of electrons and holes in the degradation process was more clearly determined by comparing charge pumping results after degradation in transistors with different offsets, although the offset transistors were limited in their ability to generate hot holes. The information gathered from all the different measurement techniques also showed that electron trapping was taking place both in the hot hole and hot electron degradation cases. From the results of energy distribution of interface states, together with gated diode results, there was evidence that the density of interface states with energies near the midgap decreased after hot electron degradation, despite the increase in the average density of interface states; a shift in the energy of interface states from deep to shallower levels was then suggested as a possible explanation.

The original intention of this project was to include LDD transistors in the offset gate experiment. However, the first technique employed for the LDD fabrication, the LDD definition by photolithography, proved to be impossible to calibrate. The second attempt, which used angled implants, was frustrated by the impossibility of removing the implant mask.

Considering the unusual characteristics of the structures used in this work, all tests were confirmed by at least one additional measurement technique even during the fabrication phase. For example, in the masking wall fabrication, the process for total oxidation of polysilicon was monitored by spectrophotometry, by measuring SiO_2 etching times and by stylus profiling. The use of multiple techniques continued through the initial assessment of the photolithographically-defined LDDs by using six different methods for the same purpose, the determination of the LDD lengths: the channel conduction

method (with different drain voltages), the paired V_{GS} method, the measurement of the gated diode currents, the detection of punchthrough by capacitance measurements and the measurement of quasi-static channel capacitance (with different values of V_{GS}). Finally, in the degradation tests, charge pumping was always used together with gated diode measurements.

The measurement of subthreshold currents for the purpose of calibrating the gate offsets had the disadvantage of being a comparative technique. In view of the inconsistencies in gate offsets introduced by distortion/rotation during the fabrication process, it would have been preferable to make use of a technique which could measure gate offsets individually. Nevertheless, the resolution of the subthreshold technique was good enough for detecting the 0.05 μm gap increments.

7.2. Suggestions for future work

Many interesting topics had to be left behind in the present work, particularly considering that the design and fabrication of the offset-gate transistors (with which some time had to be spent) was not the main objective of this project. They are left as a suggestion for future work. Additionally, it is important to let investigators who are planning to use similar techniques know of some of the difficulties encountered and how they can be avoided.

7.2.1. Device design and fabrication

Most of the distortion/rotation problems described in Chapter 4 can be minimised by positioning the offset transistors near the chip center and keeping transistors of the same array as close as possible from each other. If the transistors are close together, it might be possible to use special test structures for the calibration of both the gate offset and the LDD length. In the second chip design, the inclusion of several structures for registration measurement in different places within the chip proved to be invaluable and is strongly recommended.

The angled implant process for the production of systematic LDD increments is a promising technique because it does not require a subsequent calibration. From the

experience acquired during this project, it was concluded that the main problem is to obtain a material which is good as a mask and easy to remove without causing damage to the underlying silicon. In future experiments, it is important to attempt to leave the silicon surface protected all the time even if the chemical reaction used for the barrier removal does not etch silicon. One should bear in mind that the surface is not only heavily doped but also damaged by the ion implants, therefore exhibiting different properties than undamaged monocrystalline silicon.

An investigation on the polysilicon recrystallisation by ion implantation processes is advisable to examine the damage caused to the underlying SiO_2. As mentioned in Chapter 4, this damage was optically visible after the second gate patterning during the processing of the first batch. This effect is not usually observed because in conventional processes polysilicon is quite often doped using a solid source in a diffusion furnace instead of ion implantation. It is therefore suggested that capacitors and transistors be fabricated using polysilicon gates deposited at different temperatures and implanted with different energies and doses. This will result, after recrystallisation, in crystallites of different sizes. A subsequent test of the interface properties would allow a correlation between the recrystallisation and the interface quality.

7.2.2. Device testing

The results obtained in this work could be complemented by the use of larger transistors, which would allow the use of accurate capacitance measurements. It would be particularly interesting to confirm the degradation and interface state generation models by carrying out DLTS (deep-level transient spectroscopy) and quasi-static C-V tests. Because some of the parameters tested in the present work (e.g. the charge pumping current) scale up as the device size is increased, the use of larger transistors would also allow the confirmation of some of the phenomena observed here.

The possibility of producing LDDs with systematically incremented lengths constitutes not only a powerful tool for the study of hot carrier phenomena but is also an invaluable aid for the analysis of the next generation of submicron IGFETs. The photolithographic method for obtaining LDD transistor arrays is certainly feasible. The only reasons why it had to be substituted by the angled implant technique in the present study

were the field distortions observed during photolithography and the inability of calibrating both the LDD length and the gate offset at the same time. Now that the distortions are better understood and better controlled, it is possible to repeat the LDD experiment with better chances of obtaining consistent results. In addition to the LDD calibration methods introduced in Chapter 5, the relevance of the information to be extracted from such an experiment certainly justifies pursuing new calibration alternatives.

It is also suggested that the experiments are repeated for p-type transistors. This would be particularly important for the observation of hot hole degradation in such devices. For both n and p-type transistors, it is recommended the use of a wider range of bias conditions. Low temperature (77 K) stress, which in this work was used simply to confirm the nature of degradation, also requires further investigation.