The physics and technology of submicron MOS devices

Citation for published version (APA):
Walker, A. J. (1994). *The physics and technology of submicron MOS devices*. [Phd Thesis 2 (Research NOT TU/e / Graduation TU/e), Electrical Engineering]. Technische Universiteit Eindhoven.
https://doi.org/10.6100/IR412165

DOI:
10.6100/IR412165

Document status and date:
Published: 01/01/1994

Document Version:
Publisher’s PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:

• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license above, please follow below link for the End User Agreement:

www.tue.nl/taverne

Take down policy
If you believe that this document breaches copyright please contact us at:

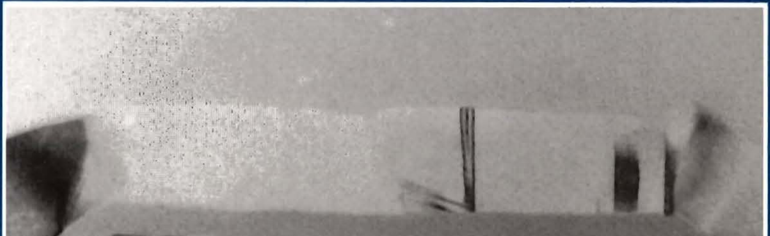
openaccess@tue.nl

providing details and we will investigate your claim.

Download date: 04. Oct. 2023

The Physics and Technology of Submicron Mos Devices

A.J. Walker



THE PHYSICS AND TECHNOLOGY OF SUBMICRON MOS DEVICES

The figure on the front cover shows a TEM cross section of an Electrically Programmable Read Only Memory cell obtained using a focussed ion beam to slice through the sample.

CIP gegeven:
ISBN: 90-74445-08-X

THE PHYSICS AND TECHNOLOGY OF SUBMICRON MOS DEVICES

PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Technische Universiteit Eindhoven,
op gezag van de Rector Magnificus,
prof.dr. J.H. van Lint,
voor een commissie
aangewezen door het College van Dekanen,
in het openbaar te verdedigen
op dinsdag 8 februari 1994 te 16.00 uur

door

Andrew Jan Walker

geboren te Sialkot, Pakistan

Dit proefschrift is goedgekeurd door de promotoren:

prof.dr. F.M. Klaassen en prof.dr. P.H. Woerlee

The work described in this thesis has been carried out under supervision of prof.dr. F.M. Klaassen at the Philips Research Laboratories in Eindhoven, The Netherlands, as part of the Philips Research programme.

To my cousin Gerda

Contents

Chapter 1

General Introduction

"It seems as if a great deal were attainable in a world where there are so many marriages and decisive battles, and where we all, at certain hours of the day, and with great gusto and despatch, stow a portion of victuals finally and irretrievably into the bag which contains us."
R.L. Stevenson

1.1 Introduction

J.E. Lilienfeld could not have known what he was starting when he filed the first patents on amplifying devices using the field effect in 1926 and 1928 [1]. Leipzig of the 1920s, where he was professor, seems as far removed from the present day as the first MOSFETs (metal-oxide-semiconductor field-effect transistor) are from the devices now on offer. The first successful MOSFET was reported in 1960 by D. Kahng and M.M. Atalla [2] who made use of thermally grown silicon dioxide as gate insulation. As in most fields of human endeavour, one parameter is taken to differentiate the various players in the game. In this case, it is minimum feature size, that ubiquitous measure of technological prowess which has gone from above 20μm in 1960 to around 0.5μm at the present time. Such is the power and importance of the semiconductor industry that a Japanese company is willing to spend 770 million American dollars in 1992 to prepare one of its factories to make 64Mbit DRAM (dynamic random access memories) chips [3]. These chips are regarded as the drivers of the industry resulting in an annual decrease of minimum feature size of 13% since the early 1960s. Such is the inexorable march to smaller dimensions that there is already talk of a 1Gbit SRAM (static random access memory) in a 0.12μm manufacturing process by the year 2000.

Why this ceaseless march, grudging ourselves the time for rest ? The answer lies, of course, with money. By reducing the minimum feature size, the area of a transistor decreases with the result that more functionality can be packed onto a given area of silicon. Less silicon per switching element means reduced costs and therefore higher profit margins. However, there are other advantages from microminiaturisation which may be obscured by the cost analysis alone. For instance, in most cases the resultant switching element dissipates less power producing less heat. Furthermore, reliability and switching time improve. These simple facts have led to the relentless pursuit of smaller dimensions in the area of ULSI (ultra large scale integration).

1.1.1 Scaling

Quantity	**Scaling factor**
Device dimensions	$1/\kappa$
Area	$1/\kappa^2$
Packing density	κ^2
Doping concentration	κ
Voltages and threshold	$1/\kappa$
Currents	$1/\kappa$
Power dissipation per unit area	1
Capacitance per unit area	κ
Charge per unit area	1
Electric field	1
Body effect coefficient	$1/\kappa^{0.5}$
Transistor transit time	$1/\kappa$
Power delay product	$1/\kappa^3$

TABLE 1.1 : Constant Field Scaling

Reducing the gatelength of a MOSFET is the first action in the downscaling of the device. A consequence is that depletion region widths must also be reduced to prevent the source and drain depletion regions from meeting and punching through. This can be attained by increasing

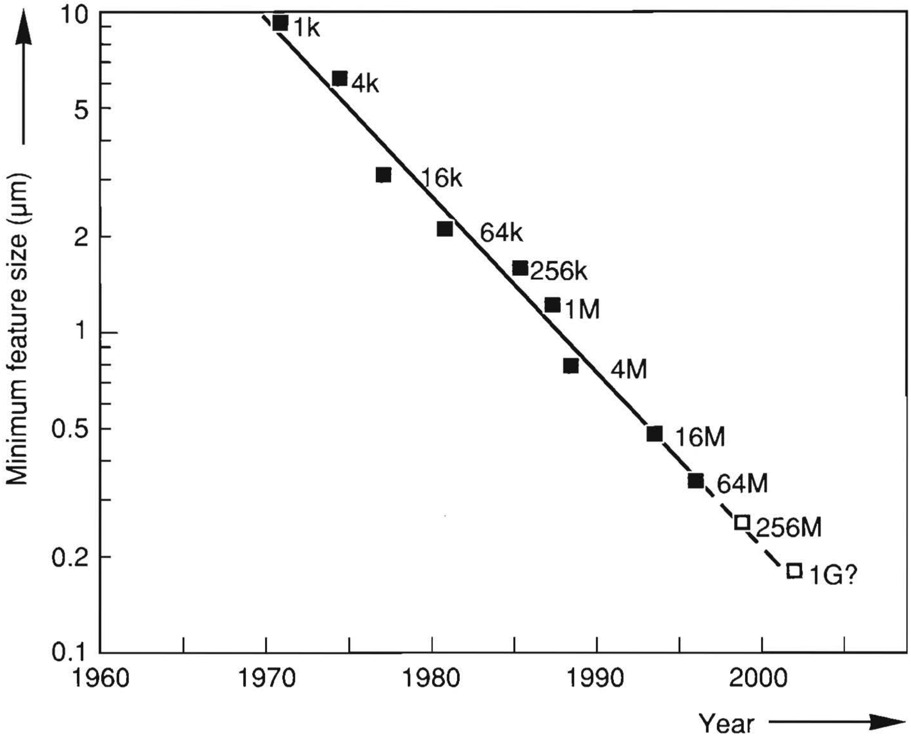


Figure 1.1. Historical evolution of the minimum feature size for dynamic random access memories at their introduction in production.

the substrate doping and decreasing the power supply voltage. The increase in doping results in threshold voltages being too high requiring, in turn, that the gate oxide thickness also be reduced. In other words, the process architecture must be changed to suppress short channel effects.

The incessant march to smaller dimensions can be seen in Fig. 1.1 which shows the historical evolution of the minimum feature size (the MOS gate length) for dynamic random access memories (DRAM) at their introduction in production [4].

The first set of rules proposed for the scaling of the MOSFET was introduced in 1974 [5] and is called *constant field scaling.* In this approach, the maximum electric field and the internal electric field profile are unaltered going from the original large device to the small scaled version. Using this method, the MOSFET is reduced by a factor $1/\kappa$

(κ larger than 1) in all three dimensions (length, width, junction depth, gate oxide thickness). The effect of this is summarised in Table 1.1.

The problems with voltage losses along interconnections and in contact holes along with the fact that the slope of logI_{ds} versus V_{gs} (I_{ds} - source drain current; V_{gs} - gate source voltage) does not change [6] mean that a larger fraction of the available voltage is lost in resistances and voltage swings to turn transistors on. This has led to the development of *constant voltage scaling* in which the voltage is kept constant as the device is scaled by a factor $1/\kappa$ ($\kappa > 1$). The main differences between this approach and that of the constant field scaling is that the supply and threshold voltages remain constant and the gate oxide thickness is reduced by a smaller amount than other dimensions to avoid problems with high fields. This approach has a more practical foundation than others in that standard supply voltages have remained largely constant while transistor dimensions have shrunk. It also deals with the practical problem that the threshold voltage does not scale downwards as dimensions shrink. With a reduced supply voltage, this would mean that less voltage over and above the threshold voltage (the "gate drive") would be available to invert the channel resulting in smaller currents.

Other approaches to scaling have been suggested such as *quasi-constant voltage scaling* in which dimensions and dope levels are scaled as in the constant field approach but voltages are reduced by a smaller amount [7]. This means that depletion regions do not scale by the same amount as transistor dimensions. To achieve this, substrate doping is scaled by another factor resulting in the generalised scaling rules [8].

All the above approaches can be found in Table 1.2.

Quantity	Constant Field	Constant Voltage $1 < \beta < \kappa$	Quasi-constant Voltage $1 < \beta < \kappa$	Generalised $1 < \beta < \kappa$
Length, Width	$1/\kappa$	$1/\kappa$	$1/\kappa$	$1/\kappa$
Gate oxide	$1/\kappa$	$1/\beta$	$1/\kappa$	$1/\kappa$
Substrate dope	κ	κ	κ	κ^2/β
Voltages	$1/\kappa$	1	$1/\kappa^{0.5}$	$1/\beta$

TABLE 1.2 : Different Scaling Rules

Another scaling procedure takes advantage of the empirical result that devices exhibit long channel behaviour (i.e. short channel effects are absent) if their channel lengths are longer than some minimum amount given as

$$L_{min} = (constant)[x_j t_{ox}(d_s + d_d)^2]^{1/3}$$

where x_j is the source/drain junction depth, t_{ox} is the gate oxide thickness, and d_s and d_d are the source and drain depletion region widths [9].

1.1.2 Process Technology

The key to miniaturisation is the huge advancements made in process technology. Billions of dollars are spent in the areas of lithography, deposition and etching to enable finer dimensions to be made on silicon wafers. Optical lithography is the workhorse of the semiconductor industry with dimensions down to 0.5μm possible using the i-line of a mercury-arc light source to expose the resist on the wafer. The use of excimer laser light sources is expected to allow features down to 0.25μm [10]. Beyond that, X-ray and electron-beam lithography are possible candidates for future processes. The fabrication of semiconductor circuits requires the deposition of various materials including dielectrics, semiconductors and metals. These processes demand the deposition of ultra clean layers, from a few nanometers to over 1 micron in thickness, on large silicon wafers of 8 inches ($\approx$ 20cm) in diameter with a thickness variation across the wafer of less than 5%. The patterns defined in the resist by lithography are transferred to the wafer by etching. As devices shrink, the requirement is for anisotropic etches that remove material much faster in the vertical than in the horizontal dimension. In this way, the pattern in the resist is faithfully transferred to the underlying material without much loss in the linewidth. These anisotropic processes are achieved with the use of plasmas where reactive species are made to bombard the wafer surface. The fabrication of a ULSI circuit requires the use of hundreds of process steps involving lithography, deposition, etching, oxidation, implantation, high temperature annealing and cleaning. All for the sake of smaller devices and larger returns on investments.

1.2 Thesis Summary

The research to be described concerns topics related to the lateral and vertical scaling of devices for ULSI. In this case, the devices are metal-oxide-semiconductor (MOS) transistors and are fabricated in silicon.

In the scaling of the MOS transistor to smaller dimensions, the gate oxide thickness reduces while the substrate doping concentration increases. These steps lead to an increase in the vertical electric fields in the channel region of the device resulting in the reduction of the mobility of the charge carriers. The modeling of this mobility reduction is important to be able to simulate circuits containing submicron devices. The first main topic describes the modeling of the electron and hole inversion layer mobility as a function of the effective normal field, E_{eff} [11].

Scaling MOS transistors to smaller dimensions also involves the reduction in the source and drain junction depth. While shallow junctions are required in both NMOS and PMOS devices, the latter provide the greater challenge because of the high diffusivity of boron [12] and boron's tendency to channel during implantation [13]. In the second main topic, work will be described where the surface of silicon is made amorphous prior to boron implantation which reduces the channeling and affects the diffusivity. With the use of cross sectional Transmission Electron Microscopy (XTEM), Rutherford Backscattering (RBS) and Secondary Ion Mass Spectroscopy (SIMS), a complete description of this technique will be presented.

In the scaling of PMOS transistors, it is known that the use of the standard n-type polycrystalline silicon as gate material leads to problems. This involves the workfunction of the gate being too low and the magnitude of the threshold voltage too high, which can be circumvented by a p-type implant in the channel region of the device. The resulting vertical p-n junction in the channel makes it very difficult to shrink this device any further [14]. Increasing the gate's workfunction would help to solve this problem. This can be done using boron to dope the gate. The problems and limitations of this approach will be described in the third main topic.

In integrated circuits containing scaled devices, the voltage which can be applied safely to any part of the circuit reduces due to thinner oxides and shallower junctions. In applications where devices are integrated that require large voltages for operation, problems can arise. This is especially the case for Electrically Erasable and Programmable Read Only Memories (EEPROM) and Flash EEPROM cells. These normally require voltages above 12 volts. If these are embedded in submicron circuits, special process steps have to be carried out to enable such large voltages to be used. The fourth main topic deals with a new techique to reduce the barrier height which limits electron tunneling through a thin oxide. Therefore, voltages can be reduced to program and erase EEPROM and Flash cells treated using this technique. In this way, such devices could be integrated with scaled MOS transistors.

In summary, the thesis deals with the practical problems of scaling MOS transistors to smaller dimensions. It is divided into four main topics and is based on several publications as follows:

1 Electron and Hole Inversion Layer Mobility [15], [16].

2 Shallow Junctions [17], [18].

3 Gate Work Function Engineering [19]

4 Silicon/Silicon Dioxide Barrier Height Adjustment [20].

1.3 References

[1] J.E. Lilienfeld, U.S. Patents 1,745,175 (filed in 1926, issued in 1930), 1,877,140 (filed in 1928, issued in 1932), and 1,900,018 (filed in 1928, issued in 1933).

[2] D. Kahng and M.M. Atalla, IRE-AIEE Solid State Device Res. Conference, Carnegie Institute of Technology, Pittsburgh, PA 1960.

[3] "Falling off the learning curve", The Economist Year Book, 1992 Edition, p182, 1992.

[4] D. Bois, "The ultimate evolution of CMOS technology," Commission of the European Communities, Directorate-General III, DG III-A-5, 1993.

[5] R.H. Dennard, F.H. Gaensslen, H.N. Yu, V.L. Rideout, E. Bassous and A.R. LeBlanc, IEEE J. of Solid-State Circuits, vol. SC-9, pp. 256-268, 1974.

[6] "Operation and Modeling of the MOS Transistor," by Y. Tsividis, p204, McGraw-Hill, 1987.

[7] P.K. Chatterjee, W.R. Hunter, T.C. Holloway and Y.T. Lin, IEEE Elec. Dev. Lett., vol. EDL-1, pp. 220-223, 1980.

[8] G. Baccarani, M.R. Wordeman and R.H. Dennard, IEEE Trans. Elect. Dev., vol. ED-31, pp. 452-462, 1984.

[9] J.R. Brews, W. Fichtner, E.H. Nicolian and S.M. Sze, IEEE Elect. Dev. Lett., vol. EDL-1, pp. 2-3, 1980.

[10] E.H. Kung, M. Cheng, O. Nalamasu, A.G. Timko, C.B. Case and V. Pol, Jap. J. of Appl. Phys., vol.30, no. 11B, pp. 3030-3036, Nov. 1991.

[11] A.G. Sabnis and J.T. Clemens, IEDM Tech. Dig., pp. 18-21, Dec. 1979.

[12] T.M. Liu and W.G. Oldham, IEEE Electron Device Lett., vol. EDL-4, p.59 1983.

[13] T.O. Sedgwick, A.E. Michel, V.R. Deline, S.A. Cohen and J.B. Lasky, J. Appl. Phys. 63 (5) pp. 1452-1463, March 1987.

[14] G.J. Hu and R.H. Bruce, IEEE Trans. Elec. Dev., ED-32, pp. 584-588, March 1985.

[15] A.J. Walker and P.H. Woerlee, European Solid State Device Research Conference, Tech. Dig. pp. 667 - 670 Sept. 1987.

[16] A.J. Walker and P.H. Woerlee, J. Phys.Coll., vol. C4, pp. 265-268, Sept. 1988.

[17] A.J. Walker, J. Appl. Phys. 71 (4) pp. 2033-2035, February 1992.

[18] A.J. Walker, P.H. Woerlee, H.G. Pomp and N.E.B. Cowern, J. Appl. Phys. 73 (8) pp. 4048-4053, April 1993.

[19] A.J. Walker and P.H. Woerlee, pp.29-32, Proc. 19th European Solid State Device Research Conference, Berlin, 1989.

[20] A.J. Walker and J. Politiek, Appl. Phys. Lett., vol.63, pp.758-760, 1993.

Chapter 2

Inversion Layer Mobility

> *"And it would seem also, on a hasty view, that the attainment of as much as possible was the one goal of man's contentious life. And yet, as regards the spirit, this is but a semblance."*
>
> *R.L. Stevenson*

2.1 Introduction

The mobility of charge carriers in MOS inversion layers is one of the most fundamental parameters for MOSFET device modeling. All the most important device characteristics such as source - drain current and switching time include the mobility in their expressions. For a full analysis of the MOSFET, the reader is referred to the literature [1]. In this chapter, a simple qualitative approach is given on MOSFET operation. In addition, the scattering mechanisms that affect charge carriers in the MOS inversion layer are described. Finally, the new MOSFET inversion layer mobility model based on effective normal field, E_{eff}, is presented.

2.2 MOSFET: Qualitative Approach

The structure of the silicon MOSFET has essentially remained the same since its inception with dimensions being reduced with time. A schematic cross section of an n-type MOSFET can be seen in figure 2.1

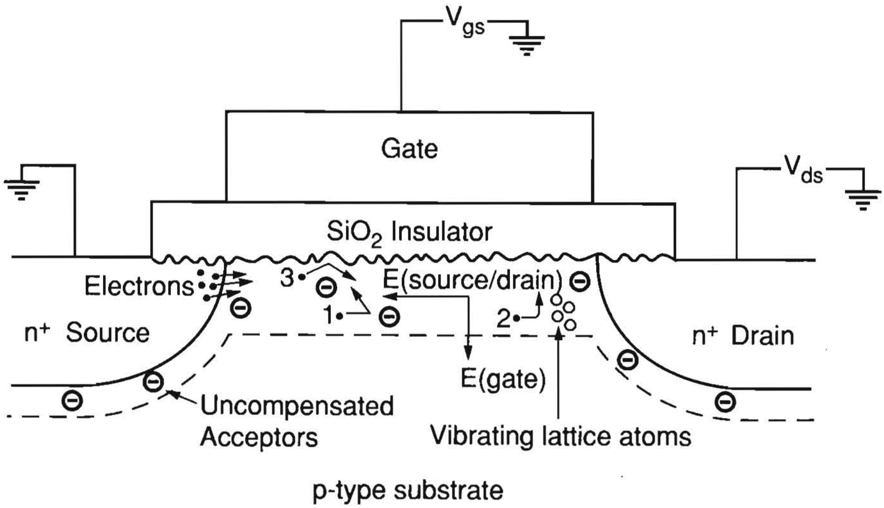


Figure 2.1. Schematic cross section of n-type MOSFET showing fields and scattering mechanisms acting on the mobile electrons in the inversion layer.

In this device, a silicon dioxide insulator separates the gate electrode from the n-type source and drain and the p-type substrate. In present day devices, the gate electrode is fabricated from polycrystalline silicon which is degenerately doped. The source and drain are self-aligned to this gate in the sense that the gate acts as a mask for source and drain implantation thus resulting in minimal gate to drain and gate to source capacitance. Also given in figure 2.1 are the fields affecting the movement of the electrons along with the scattering mechanisms that the electrons experience in the inversion layer. More information on these points is given later.

To change the band bending at the silicon - silicon dioxide (Si/SiO_2) interface, a voltage V_{GB} is applied between the gate and substrate electrodes. In this simple case, the source and substrate are both at ground potential. As V_{GB} is swept from negative values, through zero to positive values, the band bending at the Si/SiO_2 interface changes from accumulation through depletion to inversion. These three cases are given in figure 2.2. This shows the band bending at the surface of the silicon as a function of depth in the channel region of a MOS transistor. The positions of the source and drain are therefore out of the plane of the paper. With a negative voltage on the gate with respect to the source and substrate, holes in the substrate are attracted to the surface forming an

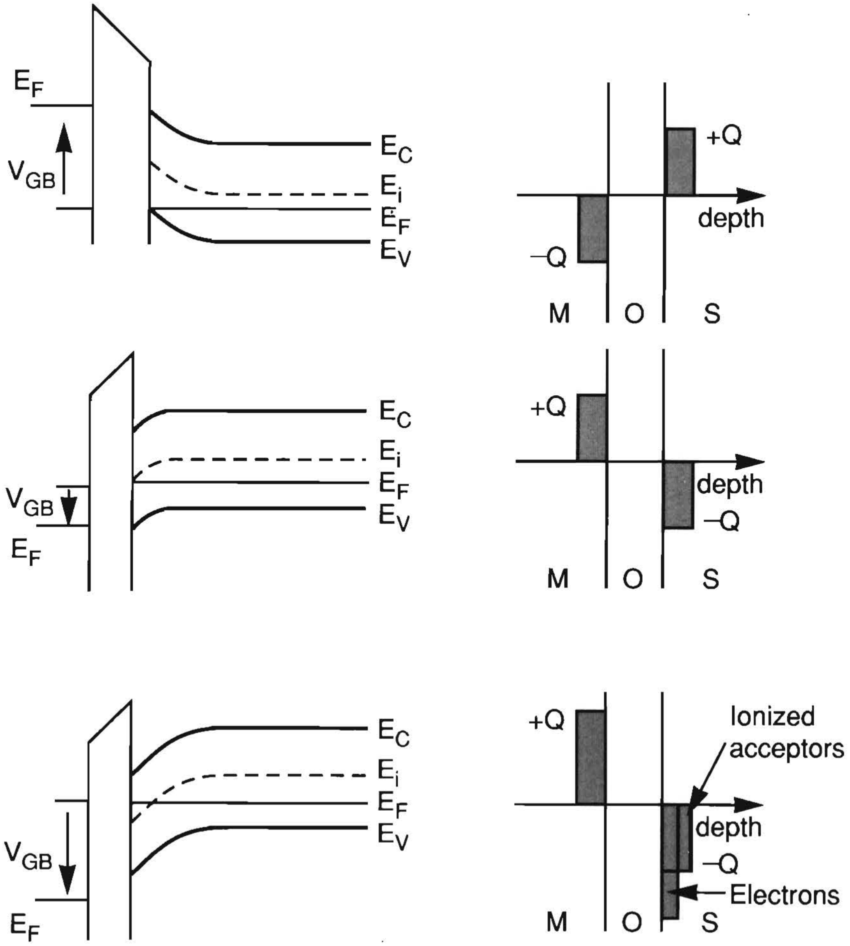


Figure 2.2. Band bending at the surface of the silicon due to the application of a voltage on the gate electrode.

accumulation layer. As V_{GB} becomes less negative, the holes are driven from this layer with the band bending becoming less as shown in figure 2.2. At the flat band voltage condition between gate and substrate, the bands are literally flat with no net charge in the substrate.

As V_{GB} sweeps through zero to positive values, the band bending is in the opposite direction to the accumulation case. In the first instance, holes are pushed from the surface layer resulting in negatively charged uncompensated acceptors being present in the depletion layer.

At still larger positive V_{GB} values, electrons are attracted to the surface coming primarily from the source electrode. The resulting inversion layer is a negatively charged conducting layer with mobile charge, Q_{inv}, due to these electrons. Since Q_{inv} is an exponential function of band bending at the silicon surface, band bending saturates as gate voltage increases. Effectively, the depletion region reaches a maximum depth into the silicon.

By applying a voltage, V_{DS}, now between source and drain, a current of electrons can be made to flow.

These are the essentials of the MOSFET.

2.3 Scattering Mechanisms

Charge carriers travelling from source to drain in a silicon inversion layer experience scattering of their motion due to various causes as is shown in figure 2.1. First, the uncompensated acceptors (for nMOS devices) or the uncompensated donors (for pMOS devices) are of the same charge type as the mobile inversion charge leading to Coulomb repulsion. This results in scattering. As the gate voltage increases leading to increased inversion, more mobile charge carriers are attracted into the channel region. This in turn leads to screening of the effect of the immobile donors or acceptors. Therefore, Coulomb scattering is important at low gate voltages and at high channel doping levels. Another more important source of Coulomb scattering is charges in interface states and in the gate oxide.

Second, bulk phonons and surface phonons (surfons) from the quantum vibrations of the crystal lattice scatter the mobile charge carriers.

Third, the interface between the crystal silicon and the gate oxide is not atomically smooth. This interface roughness scatters the mobile charge carriers. This effect becomes more important as the charge carriers are forced closer to this interface by increasing the gate voltage to increase the mobile inversion charge.

The net effect of all this scattering is seen in the effective mobility of the charge carriers moving from source to drain in the MOS inversion layer.

2.4 Inversion Layer Mobility

The scaling of MOS transistors towards the requisite submicron dimensions involves the use of thinner gate oxides, higher substrate doping and/or lower operating voltages. The goal of these changes is to increase the performance of the individual devices, but such changes may compromise the very performance they seek to enhance. Performance degradation can be seen by analysing the mobility of the charge carriers in the silicon inversion layer and how this is affected by gate oxide thickness, surface doping density and gate voltage. Figure 2.3 shows mobility degradation with gate voltage for scaled NMOS transistors with different gate oxide thicknesses and surface doping densities. The ability to model such effects is of great importance for device simulations which are necessary to direct actual experiments in silicon and thus reduce expenditure on these experiments.

As carriers travel from source to drain in an MOS transistor, they experience an electric field made up of two components. Firstly, a lateral contribultion due to the source - drain voltage and, secondly, a perpendicular component due to the gate voltage. Each component affects the carrier mobility in different ways. These electric fields are indicated for an NMOS in figure 2.1 in schematic detail.

This chapter presents a physically - based model for the perpendicular effective field dependence of the electron and hole mobilities in silicon inversion layers at low lateral fields. Recent results of an investigation into the application of this model to measurements carried out on MOS devices with gate oxide thicknesses and surface doping densities scaled for process generations down to 0.5μm are given. Furthermore, for the first time such a model was incorporated in a two - dimensional device simulation program and comparisons between predictions from this program and measurements on submicron transistors are presented.

2.4.1 Model

The modeling of the inversion layer mobility for applications in device simulation seems well established as evidenced by models from Yamaguchi [2] and Hiroki et al. [3]. Generally the local normal electric field is used to evaluate the influence of the normal field on the mobility. Such an approach is very attractive for applications in a device simulator. However, from a physical point of view it has several

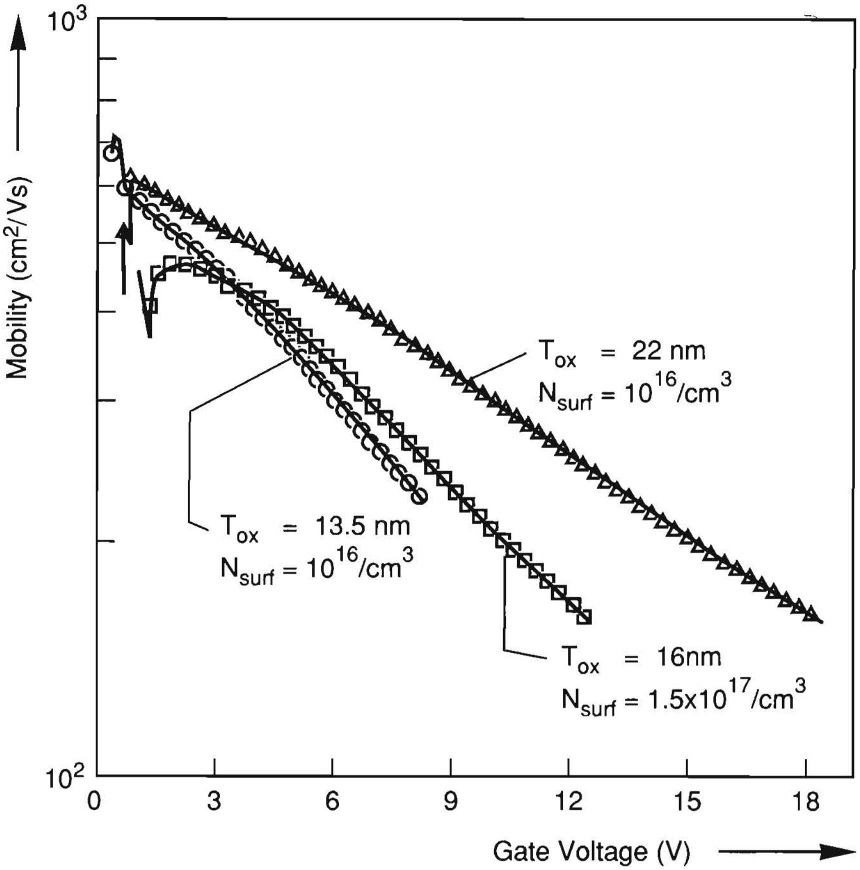


Figure 2.3. Mobility degradation as a function of gate voltage for various n-type MOSFETs with different gate oxide thicknesses and surface doping densities. The source/drain voltage was 100 mV.

drawbacks. Firstly, since the normal field can vary strongly over the inversion layer thickness, appreciable variations of the mobility can occur over distances significantly smaller than the electron mean free path. More importantly, it has been shown elsewhere that the electron [4] and hole [5] mobilities follow universal curves when plotted as functions of an effective normal field, E_{eff}, given by

$$E_{eff} = \frac{1}{\epsilon_{si}}(Q_{dep} + \eta Q_{inv}) \tag{2.1}$$

with $\eta = 1/2$ for electrons and $1/3$ for holes and where Q_{dep} and Q_{inv} are the depletion and inversion layer charges per unit surface area respectively. This effective normal field can be regarded as the average normal field over the depth of the inversion layer.

In this new model a semi - empirical approach was taken to model the mobility - E_{eff} curves. It was assumed that the room temperature inversion layer mobility is dominated by three scattering mechanisms [6]:

1 Coulomb scattering

2 Carrier - phonon scattering

3 Surface roughness scattering

Each has its own contribution to the net mobility. These contributions are designated by μ_c, μ_{cp} and μ_{sr} respectively. Actual modelling of scattering processes in inversion layers is very complex due to the quantum mechanical nature of these processes and the fact that at temperatures above absolute zero more than one sub - band is filled. Therefore, a simplified semi - empirical approach was adopted as a pragmatic step towards obtaining a fitting model. The three scattering processes are schematically shown in figure 2.1.

To obtain a simple first order expression for carrier mobility due to screened Coulomb scattering, the Brooks - Herring formula was adopted [7]:

$$\mu_c = \frac{CT^{1.5}}{ln(1+b) - b/(1+b)} \cdot \frac{1}{N_I} \tag{2.2}$$

where $b = (24m^* \epsilon_{Si} k^2 T^2)/(\hbar^2 q^2 N_o)$, N_I is the charged impurity density, N_o is the mobile carrier density, C is a constant, ϵ_{Si} is the dielectric constant of silicon and the other symbols have their usual meanings.

However, the following modifications were made for the inversion layer :

$$N_o = n_{inv} / < z > \tag{2.3}$$

$$N_I = N_I^1 + N_{ox} / < z > \tag{2.4}$$

where n_{inv} is the number of mobile carriers per unit surface area, $< z >$ is the average distance of mobile carriers from the Si/SiO_2 interface, N_{ox} is the fixed oxide charge per unit surface area and N_I^1 is the actual charged impurity density due to uncompensated donors or acceptors.

As a result of numerical calculations [8], $< z >$ can be given as a function of the mobile inversion charge number density per unit surface area, n_{inv}, and therefore as a funtion of E_{eff}.

For carrier - phonon scattering, the prediction from theory (see [9] and references therein) is

$$\mu_{cp} \approx aT^{-1}(Q_{inv}/3 + Q_{dep})^{-\frac{1}{3}} \tag{2.5}$$

Surface roughness scattering, according to theory (see references in [9]), obeys the following relation :

$$\mu_{sr} = b(Q_{inv}/2 + Q_{dep})^{-2} \tag{2.6}$$

These two equations for μ_{cp} and μ_{sr} can be reformulated as functions of the effective perpendicular field, E_{eff} by considering Gauss's Law in one dimension. Reference is made to figure 2.4 which shows electron potential plotted as a function of depth into the silicon. Using Gauss's Law at the surface of the inversion layer, it can be seen that the field at the surface is

$$E_{surf} = \frac{1}{\epsilon_s}(Q_{inv} + Q_{dep}) \tag{2.7}$$

while at the inversion layer/depletion layer interface, it is

$$E_{dep} = \frac{Q_{dep}}{\epsilon_s} \tag{2.8}$$

These are of course the gradients of the curve in figure 2.4 at the respective depths. Therefore, the average of these two fields is defined as the effective normal field, E_{eff}, and is given as

$$E_{eff} = \frac{1}{\epsilon_s}(\frac{Q_{inv}}{2} + Q_{dep}) \tag{2.9}$$

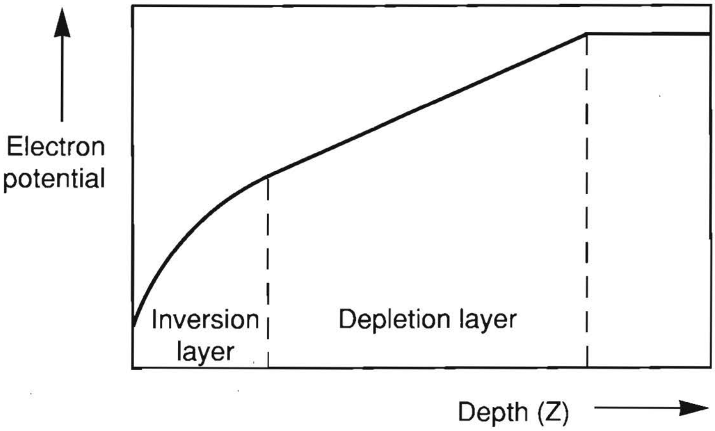


Figure 2.4. Electron potential as a function of depth from the silicon/silicon dioxide interface through the inversion and depletion layers.

Therefore, we have

$$\mu_{cp} = k_{cp} T^{-1} E_{eff}^{-1/3} \tag{2.10}$$

and

$$\mu_{sr} = k_{sr} E_{eff}^{-2} \tag{2.11}$$

where k_{cp} and k_{sr} are constants.
The net mobility μ is calculated using Matthiesen's rule. i.e.

$$\frac{1}{\mu} = \frac{1}{\mu_c} + \frac{1}{\mu_{cp}} + \frac{1}{\mu_{sr}} \tag{2.12}$$

with μ_c, μ_{cp} and μ_{sr} given in equations 2, 5 and 6 respectively with the constants C, k_{cp} and k_{sr} as adjustable parameters.

The above equations were incorporated into a two - dimensional device simulation program, CURRY [10], in which the effective field at any point from source to drain was calculated by averaging the normal field in the depth direction over the inversion layer or surface accumulation layer at that point. In other words, the following approximation is used for the calculation of the effective field :

$$E_{eff}(x) = \int_0^d n(x,y) E_y(x,y) dy / Q_{inv}(x) \tag{2.13}$$

where d is the thickness of the inversion layer and x is the lateral coordinate from source to drain. The model is therefore a non - local one with the mobility at any point dependent on the values of the normal electric field at other points. This contrasts with the approach used in some device simulators where the effective mobility at the point (x,y) is determined by the value of the normal electric field, $E_y(x, y)$, at that point.

2.4.2 Mobility Measurement

In order to exclude possible two- dimensional effects playing an important role, the mobility at small lateral source - drain fields was extracted from measurements on large MOS transistors. If a small source - drain voltage, V_{ds}, is applied to such a device and the gate - source voltage, V_{gs}, is varied (the substrate is shorted to the source), then a certain source - drain current, I_{ds} will flow. In this case, the low lateral field mobility is given as

$$\mu = \frac{I_{ds}}{\frac{W}{L} V_{ds} Q_{inv}} \tag{2.14}$$

where I_{ds} is the current from source to drain, W and L are the width and length of the transistor respectively, V_{ds} is the source - drain voltage, and Q_{inv} is the mobile inversion charge per unit surface area.

In most approaches, the mobile inversion charge, Q_{inv}, is approximated by

$$Q_{inv} = C_{ox}(V_{gs} - V_T) \tag{2.15}$$

where C_{ox} is the transistor's oxide capacitance per unit surface area, V_{gs} is as above and V_T is the transistor's threshold voltage [11]. This expression is fairly accurate at gate - source voltages much larger than the threshold voltage. However, at gate - source voltages near threshold, uncertainty arises. This is due to the fact that this expression assumes, in contrast to reality, that there are no mobile charges at the surface at voltages below threshold. This approximation can be removed by actually measuring Q_{inv} on capacitor structures [12]. In addition, the measurement of the immobile depletion charge per unit surface area, Q_{dep}, can also be determined using capacitor structures.

Since Q_{inv} and Q_{dep} are crucial to the model, a short description of how these are measured follows.

Q_{inv} and Q_{dep} are determined using the so-called "split" capacitance - voltage method [13]. A schematic representation of the experimental set-up for this measurement in the case of electron mobility is given in figure 2.5. The main point to note is that separate measurements are made of the source-drain and bulk contributions to the gate capacitance as a function of gate voltage. The d.c. gate voltage, V_G, is ramped from a negative value through threshold to a positive value so that the silicon surface changes from being strongly accumulated with holes to being strongly inverted with electrons having passed through depletion on the way.

Superimposed on the d.c. gate voltage is a small a.c. signal, dV_G, which produces the displacement currents, I_1 and I_2 shown in figure 2.5. When the silicon surface is in strong accumulation, the density of holes at and near the Si/SiO_2 interface is very large and any small signal variation in this density due to the small gate signal, dV_G, is recorded as a change in I_1. That is, the holes are considered to come exclusively from the substrate, I_1 being the hole current. In strong inversion, electrons are the majority carriers at the surface and any variation in their density due to dV_G is recorded as a change in I_2. That is, the electrons are considered to come exclusively from the n^+ regions bounding the channel. By making the assumption that, throughout the gate voltage sweep, electrons are supplied to the interface region exclusively by the n^+ regions, and holes by the substrate, these separate contributions to the gate capacitance can be independently monitored, as can be seen in the graph in figure 2.5. This assumption can be expressed as follows:

$$I_1 = \frac{dQ_{dep}}{dt} \quad and \quad I_2 = \frac{dQ_{inv}}{dt}$$

which, of course, means:

$$I_1 = \frac{dQ_{dep}}{dV_G}\frac{dV_G}{dt} \quad and \quad I_2 = \frac{dQ_{inv}}{dV_G}\frac{dV_G}{dt}$$

Since I_1 and I_2 are measurable quantities and dV_G/dt is known, the quantities dQ_{dep}/dV_G and dQ_{inv}/dV_G can be calculated. The latter two quantities are, in effect, the hole and electron contributions respectively to the small-signal capacitance of the gate-earth system. As can be seen

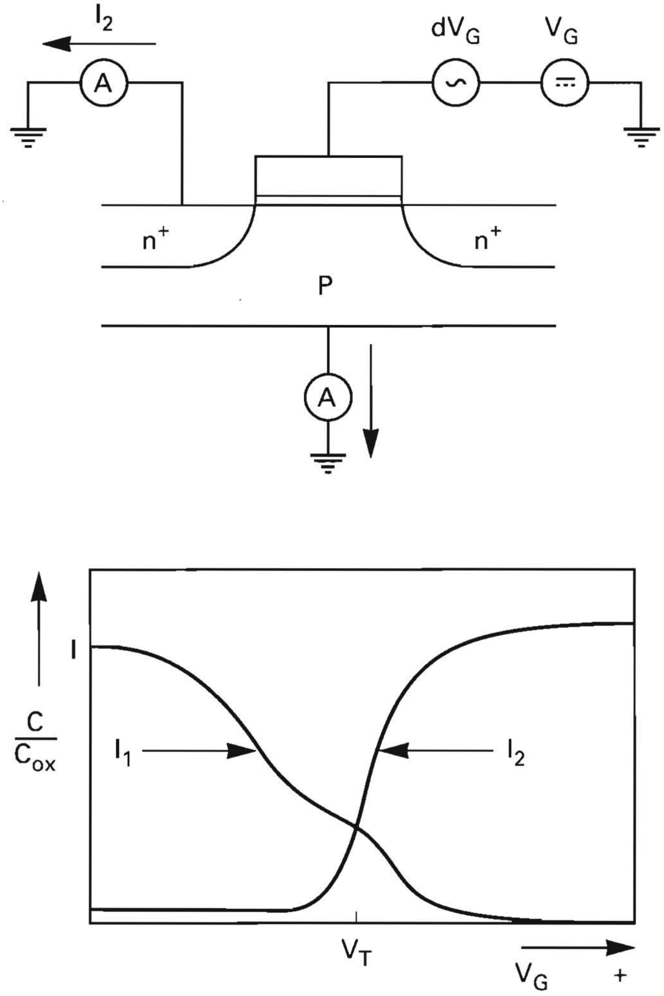


Figure 2.5. Schematic representation of the experimental set-up to measure electron mobility.

from the graph in figure 2.5, I_2 has a constant non-zero value, even in the accumulation region. This is due to the fact that the n^+ regions overlap the gate due to diffusion. This produces a parasitic capacitance since the electrons in these regions can faithfully follow a gate signal even when the channel is in accumulation, thus giving rise to a non-zero I_2.

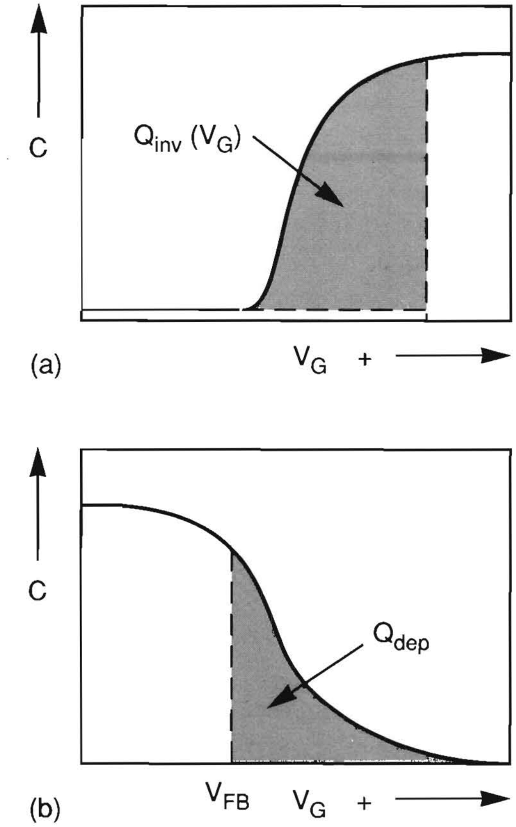


Figure 2.6. Measurement of the inversion and depletion charges from the measured capacitance-voltage curves.

Figure 2.6 shows how Q_{inv} is found as a function of V_G. From monitoring I_2, the capacitance due to the electron contribution can be plotted as a function of V_G. The background parasitic capacitance, mentioned above, is then subtracted and the resulting curve can be integrated to give $Q_{inv}(V_G)$.

After the silicon surface has reached strong inversion, the depletion charge per unit surface area, Q_{dep}, can be considered to have reached its maximum value which remains constant in inversion. As indicated in figure 2.6, the hole contribution to the total capacitance per unit surface area is integrated from the flat-band voltage, V_{FB}, to large positive V_G to give Q_{dep}.

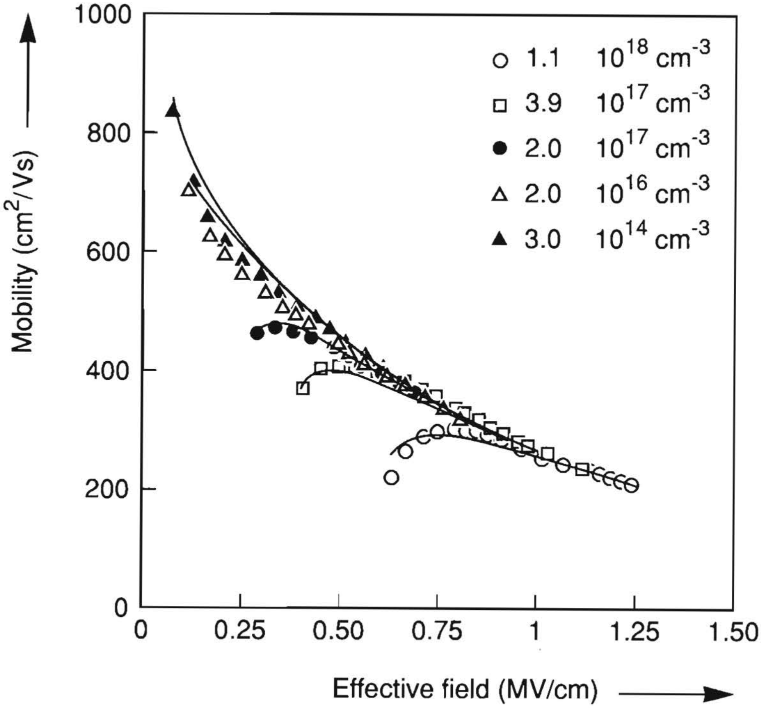


Figure 2.7. Electron mobility as a function of effective field for various surface doping concentrations with the gate oxide thickness at 14 nm for all samples.

Using the above techniques, the mobility can be directly measured as a function of gate voltage on the MOSFET. However, Q_{inv} and Q_{dep} are also known as functions of gate voltage from the capacitance measurements. Therefore, the mobility can be expressed as a function of these charges and as a function of the effective normal field, E_{eff} as required.

2.4.3 Results

The results from mobility measurements on the large n- and p-MOS devices are given in figures 2.7 and 2.8 respectively in which mobility is plotted as a function of normal effective field.

The surface doping density, N_s, and gate oxide thickness, t_{ox} are also shown in the figures. The points are the experimental data with the solid lines being the fits of the model to these data. Figure 2.7 shows

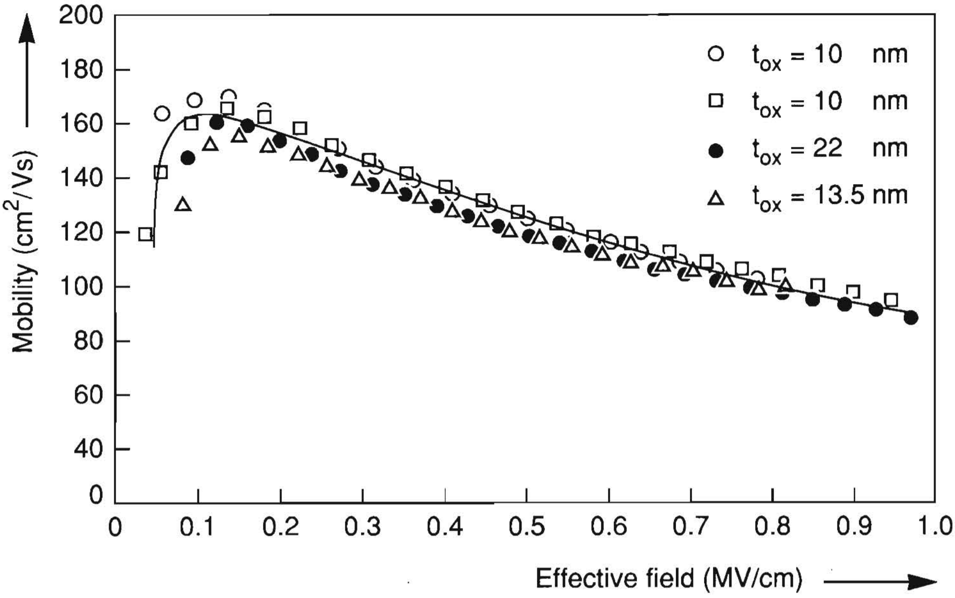


Figure 2.8. Hole mobility as a function of effective field for various gate oxide thicknesses.

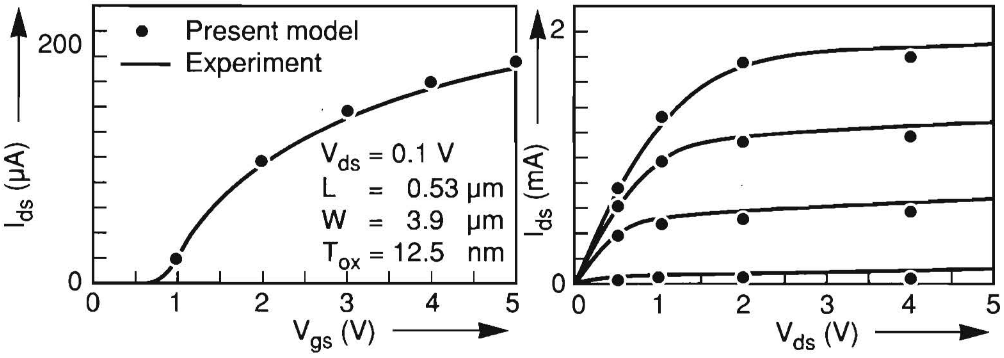


Figure 2.9. Simulations and experimental data of half-micron n-MOS transistors.

that the electron mobility in the inversion channel is almost independent of surface doping density at high effective fields but depends strongly on the doping at lower fields. In the case of the pMOS devices, different boron implants had been introduced into the originally n-type substrate but it was not clear as to what the actual density of charged impurities would be since the impurities are not completely ionized.

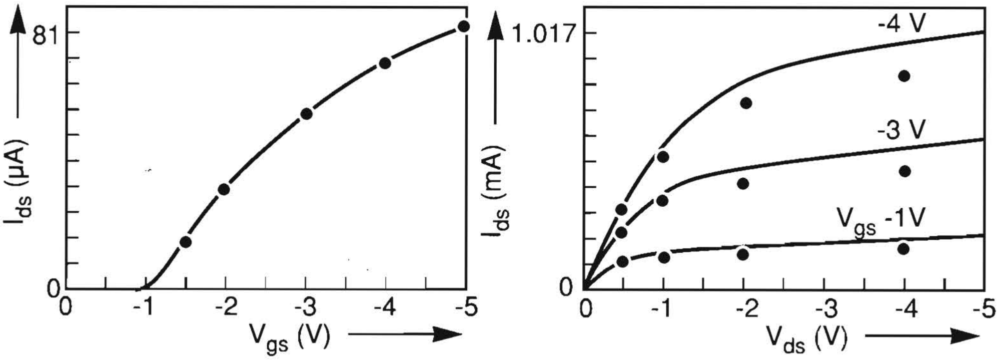


Figure 2.10. Simulations and experimental data of half-micron p-MOS transistors.

It was apparent in the model for electrons that the same values for C (equ.2), k_{cp} (equ.10) and k_{sr} (equ.11) could be used for all samples provided that the surface doping density, N_I, and the oxide fixed charge, N_{ox}, were known. The values used in figure 2.7 were $C = 2.63 \times 10^{17}$, $k_{cp} = 1.35 \times 10^7$ and $k_{sr} = 7.0 \times 10^{14}$ all expressed in cgs units.

In the model for holes, the value of N_I was unknown. Therefore a fit was carried out to find the optimum value of $C/(N_I + N_{ox}/ < z >)$ using fixed values of k_{cp} and k_{sr}. This was 0.225 for $k_{cp} = 3 \times 10^6$ and $k_{sr} = 2.22 \times 10^{14}$ all in cgs units. This is the solid line in figure 2.8

The mobility model as described was tested by modeling transistor charcateristics. Figure 2.9 shows simulations of a half-micron n-MOS transistor together with experimental data.

Good agreement between the calculated and experimental data is observed. Similar data for a buried channel p-MOS transistor are shown in figure 2.10. The $I_{ds} - V_{gs}$ characteristics at low V_{ds} are modelled well but systematic deviations in the $I_{ds} - V_{ds}$ characteristics at high drain bias suggest that the lateral - field model for holes is not sufficiently accurate at high lateral fields. The mobility model presented above is based on measurements on very large transistors. Clearly, application of the model to submicron devices leads to deviations which are not taken into account in the model itself.

2.5 Discussion and Conclusions

With the simple three parameter mobility model presented above, a large amount of experimental data can be modeled. It's incorporation in a device simulator has led to predictions of 0.5μm MOS transistor behaviour in reasonable agreement with experiment. As was seen, the modelling of hole mobility and its subsequent use in simulating PMOS current-voltage curves were less accurate than for the electron mobility and NMOS case. Future work should concentrate on hole mobility and, in particular, on the "surface channel device" where the gate polysilicon has been doped p-type by incorporation of boron. This will become the device of choice for PMOS transistors in the deep submicron regime due to the ease in making low threshold devices for low power applications (see Chapter 4). The model should be tested against the newest devices in the laboratory which have gate lengths below 0.25μm.

In the time between the appearance of the original articles [14], [15], and this dissertation, many articles have appeared along similar lines as proposed here. [16], [17], [18], [19], [20], [21], [22].

The newest most comprehensive approach can be found in [23] where an identical approach was taken regarding the effective inversion layer mobility as a Matthiesen's summation of three component mobilities due to screened Coulomb scattering, carrier-phonon scattering and surface roughness scattering. Reference was also made by the authors to the original articles by Walker and Woerlee. This confirms the approach taken here. A paper examining the effects of high substrate doping on threshold voltage and inversion layer mobility of deep-submicron MOSFETs using the original approach in [15] was published recently [24].

2.6 References

[1] "Physics of Semiconductor Devices", by S.M. Sze, 2nd ed., Wiley-Interscience, New York 1981.

[2] K. Yamaguchi, IEEE Trans. Electron Devices, vol.ED-26, pp. 1068-1074, Jul. 1979.

[3] A. Hiroki, S. Odanaka, K. Ohe and H. Esaki, IEEE Electron Device Lett., Vol.EDL-8, pp. 231-233, May 1987.

[4] A.G. Sabnis and J.T. Clemens, IEDM Tech. Dig., pp. 18-21, Dec.1979.

[5] J.T. Watt and J.D. Plummer, Symp. on VLSI tech., pp. 81-82, 1987.

[6] Y.C. Cheng and E.A. Sullivan, J. Appl.Phys., Vol.44, No.8, pp.3619 - 3625, Aug. 1973.

[7] T. Sugano, K. Hoh, H. Sakaki, T. Iizuka, K. Hirai, K. Kuroiwa and K. Kakemoto, in J. of Faculty of Engineering, Univ. of Tokyo (B) XXXII, 155, Appendix F, pp. 362-363 1973.

[8] J.A. Pals, Ph.D Thesis, Technical Univ. of Eindhoven, Dec. 1972.

[9] T. Ando, A.B. Fowler and F. Stern, Rev. of Modern Physics, 54, p.437 1982.

[10] S.J.Polak, C. den Heijer, W.H.A. Schilders and P. Markowich, in Int. J. for Num. Meth. in Eng., Vol.24, pp. 763-838, April 1987.

[11] "Device Electronics for Integrated Circuits", by R.S. Muller and T.I. Kamins, John Wiley and Sons, 1977.

[12] C.G. Sodini, T.W. Ekstedt and J.L. Moll, Solid-State Electronics, vol.25, pp. 833-841, Sept. 1982.

[13] J. Koomen, Solid State Electronics, 16, p.801, 1973.

[14] A.J. Walker and P.H. Woerlee, ESSDERC Tech. Dig. pp. 667 - 670 Sept. 1987.

[15] A.J. Walker and P.H. Woerlee, J. Phys.Coll., vol. C4, pp. 265-268, Sept. 1988.

[16] W. Hansch, R. Kircher, T. Vogelsang and M. Orlowski, Solid-State Elect., vol.32, pp.839-849, Oct. 1989.

[17] A. Hairapetian, C.R. Viswanathan and D. Gitlin, IEEE Trans. Elect. Dev., ED-36, pp.1448-1455, Aug. 1989.

[18] S. Selberherr, W. Hansch, M. Seavey and J. Slotboom, Solid-State Elect., vol.33, pp.1425-1436, Nov. 1990.

[19] M. Schubert, B. Hofflinger and R.P. Zingg, IEEE Elect. Dev. Lett., EDL-12, pp.489-491, Sept. 1991.

[20] V.M. Agostinelli, G.M. Yeric and A.F. Tasch, IEEE Trans. Computer-Aided Design of Integrated Circuits and Systems, vol.12, pp.439-445, March 1993.

[21] C.A. Dimitriadis, J. Stoemenos, P.A. Coxon, S. Friligkos, J. Antonopoulos and N.A. Economou, J. Appl. Phys., vol.73, pp. 8402-8411, 1993.

[22] H.S. Wong, Solid-State Elect., vol.36, pp. 179-188, Feb. 1993.

[23] V.M. Agostinelli, H. Shin and A.F. Tasch, IEEE Trans. Elec. Dev., vol.ED-38, pp. 151-159, Jan. 1991.

[24] M.J. van Dort, P.H. Woerlee, A.J. Walker, C.A.H. Juffermans and H. Lifka, IEEE Trans. Elec. Dev., vol.ED-39, pp. 932-938, April 1992.

Chapter 3

Shallow Implanted Boron Junctions

> *"There is always a new horizon for onward-looking men, and although we dwell on a small planet, immersed in petty business and not enduring beyond a brief period of years, we are so constituted that our hopes are inaccessible, like stars, and the term of hoping is prolonged until the term of life."*
> *R.L. Stevenson*

3.1 Introduction

Ion implantation has replaced solid-state diffusion as the preferred method of introducing dopants into silicon. A very accurately controlled amount of dopant can be introduced in this way. Since the implantation is carried out at crystal temperatures where dopant diffusion is negligible, very shallow profiles close to the crystal surface are possible. With the continued miniaturisation of silicon devices and the demand for tight control in electrical parameters, ion implantation has become an indispensable tool in very large scale integrated circuit manufacture.

The depth penetration of the ions is determined by the energy of the ions and the rate at which they lose energy to the crystal lattice. This loss is caused by electronic and nuclear scattering. At high ion velocities, electronic interactions are predominant with very few nuclear - ion interactions. As the ions slow down, nuclear scattering takes over as the main mechanism. This results in displacements of silicon atoms from their lattice sites. In other words, most silicon displacement takes place at the ion's end of range. This illustrates one of the problems of ion implantation. The method introduces defects into the silicon crytal since as little as 30 eV is sufficient to displace a silicon atom from its lattice site. These resultant silicon "self-interstitials" are a problem on two counts. First, they can coalesce to form extended defects which may lead to large parasitic leakage currents if they are situated in charge depletion layers. Second, they can interact with implanted dopant atoms leading to enhanced dopant diffusion. A third problem encountered with light ion implantation is known as "channeling" in which the ion is guided into various crystal directions and experiences far fewer retarding interactions thus leading to deeper profiles than desired. This chapter is concerned primarily with an investigation of these three effects.

The first section deals with the fabrication of shallow boron junctions using a technique known as "preamorphization" where the surface of the silicon crystal is intentionally made amorphous using a relatively heavy ion prior to the implantation of the boron. This suppresses the channeling of the boron. The heavy ion implantation however leads to silicon displacement which in turn can form large extended defects and can cause enhanced boron diffusion. The requirements for shallow, low leakage junctions using preamorphization will be presented.

The second section concentrates on the diffusion of boron in this amorphous layer. In particular, the boron was introduced by implantation of boron difluoride and the effect of the fluorine on the boron diffusion is shown during high temperature rapid thermal annealing.

3.2 Preamorphization

The lateral and vertical scaling of metal-oxide-semiconductor field-effect transistor devices for very large scale integrated circuits requires the fabrication of very shallow junctions to avoid short-channel effects. The greatest challenge arises in the fabrication of shallow p^+-n junctions, since these require the use of B which tends to channel during low energy implantation and has too low an atomic mass to amorphize the silicon at doses of practical interest. Channeling results in relatively deep penetration of the implanted B in the silicon substrate [1], [2]. The non-amorphous nature of the silicon substrate after B implantation requires the use of higher annealing temperatures to activate the dopant and remove the implantation damage [3] which, in turn, leads to considerable boron diffusion into the depth of the silicon substrate. Further problems, leading to a deepening of the p^+-n junction, arise due to anomalous enhanced diffusion of B during implant damage repair [4], [5], [6], [7] in which dopant diffusivities can be enhanced by several orders of magnitude during the initial stages of the thermal anneal.

These difficulties have led to the study of shallow p^+-n junction fabrication using preamorphization by heavy ion implantation. Numerous studies have been carried out using different heavy ions such as Si^+ [10], [11], Sn^+ [12], Ar^+ [13], Ga^+ [14] [15] In^+ [16], Ge^+ [17], [18], [19], [20] and Sb^+ [21]. The resulting amorphous layer close to the silicon surface suppresses the channeling of the implanted B ions. The layer can be regrown by solid phase epitaxy at temperatures as low as 550°C resulting in very little B diffusion and high electrical activity. Unfortunately, however, the regrowth generates extended defects located at the position of the previous amorphous-crystalline (α/c) interface which can result in very high junction leakage [19], [22]. The position of the junction depletion layer in relation to this defect band is therefore critical. As the reverse bias is increased, the depletion layer extends further into the lightly doped substrate than into the heavily doped p^+ region, and it is therefore advantageous to position the defects within the p^+ region. However, B diffusion in the region beyond the defect band has been seen to be enhanced [19] due to the supersaturation of silicon interstitials in this region, while diffusion between the defect band and the silicon surface is relatively little affected [7], [8]. Therefore, to satisfy leakage requirements, one of two approaches can be taken: (i) the B has to be implanted within the preamorphized silicon layer and diffused through the defect band, with minimal transient enhanced diffusion because the interstitial supersaturation in the region beyond the defect band only occurs at short times, before the dopant has entered this region; or (ii) the thermal annealing during the fabrication process must completely remove the implantation-induced defects.

Since thermal budgets are decreasing as transistors decrease in size, the latter approach is becoming difficult to maintain. Therefore, the former approach has been studied here to investigate conditions for ideal junction behaviour.

Structural and electrical properties of shallow junctions formed by implanting BF_2^+ into crystal silicon or into silicon preamorphized by either Si^+ or Ge^+ have been investigated [9]. The position of the junction has been varied by furnace annealing at 850°C for different times with the resultant leakage currents and forward bias ideality factors being measured.

The goal of this study is to show that by using relatively low-temperature furnace annealing, shallow p^+-n junctions can be made without having to remove all the crystalline defects formed by the implantation process. The resulting limits for ideal junction behaviour that this approach imposes will be presented.

3.2.1 Sample Preparation

The samples used fall into two categories defined by the BF_2^+ implant energy. Both consist of large diodes fabricated for electrical and physical characterization.

In the first diode experiment, phosphorus-doped n-type $< 100 >$ 0.7-0.9 Ω.cm wafers were used. A 600 nm oxide film was grown by wet oxidation at 1000°C and junction areas were defined using photolithography and wet etching in buffered HF solution. A proportion of the wafers were preamorphized by silicon implantation at 50keV and a dose of 1×10^{15} cm^{-2} or by germanium implantation at 120keV with a dose of 4×10^{14} cm^{-2}. The remaining wafers did not receive any preamorphizing implant. All the wafers were then implanted at normal incidence with BF_2^+ at an energy of 25keV and a dose of 1×10^{15} cm^{-2}. After removing some wafers for future analysis, a furnace anneal was carried out at 600°C for 60 minutes in nitrogen ambient. Again, some wafers were removed for analysis, while the remainder were furnace annealed at 850°C in nitrogen for different durations; 5 minutes, 30 minutes or 2 hours. Aluminum(2% silicon) was deposited. The metal contacts were photolithographically defined and wet etched in phosphoric acid. The process was completed with a forming gas anneal at 420°C for 30 minutes.

The second diode experiment was identical to the first except that the amorphization was carried out using only Ge implantation at 70keV with a dose of 4×10^{14} cm^{-2}. Furthermore, the BF_2^+ was implanted at 20keV instead of 25keV as in the first experiment.

All anneals within this study were performed by furnace annealing in nitrogen ambient, unless otherwise stated.

3.2.2 Results and Discussion

Diode 1 Experiment

Sample	Si Dose 50 keV 10^{15} cm^{-2}	Ge Dose 120 keV 10^{14} cm^{-2}	Anneal (min/C) After 600C	Junc. Depth Xj (nm)	PreAmorph Depth Xa (nm)	Leak at -5V (nA/cm2)	Ideal. Fact
1	1	-	-	75	90	2×10^6	1.67
2	1	-	5/850	110	90	28	1.04
3	1	-	30/850	130	90	0.68	1.01
4	1	-	120/850	185	90	0.11	1.02
5	-	-	-	175	-	1.1	1.01
6	-	-	5/850	210	-	0.03	1.01
7	-	-	30/850	220	-	0.08	1.01
8	-	-	120/850	230	-	0.07	1.01
9	-	4	-	80	125	2×10^4	1.70
10	-	4	5/850	120	125	125	1.07
11	-	4	30/850	130	125	3.5	1.02
12	-	4	120/850	180	125	0.05	1.03

TABLE I : The results of the first diode experiment.

Figures 3.1, 3.2 and 3.3 show secondary-ion mass spectroscopy (SIMS) results from the first diode experiment. As in most of the SIMS figures, 4 profiles are shown: (i) as-implanted; (ii) after 600^{o}C 60 minutes followed by 850^{o}C 5 minutes; (iii) after 600^{o}C 60 minutes followed by 850^{o}C 30 minutes; (iv) after 600^{o}C 60 minutes followed by 850^{o}C 2 hours. The profile after the 600^{o}C, 60 min anneal for solid phase epitaxial regrowth of the amorphous layer was indistinguishable from the as-implanted profile and has therefore been omitted from all SIMS figures except the implantation of BF_2^+ into crystalline silicon where a significant displacement was observed at concentrations below 3×10^{17} cm^{-3}.

Figure 3.1 shows the B profiles from the samples preamorphized by Si implantation. The 850^{o}C anneals push the B deeper into the

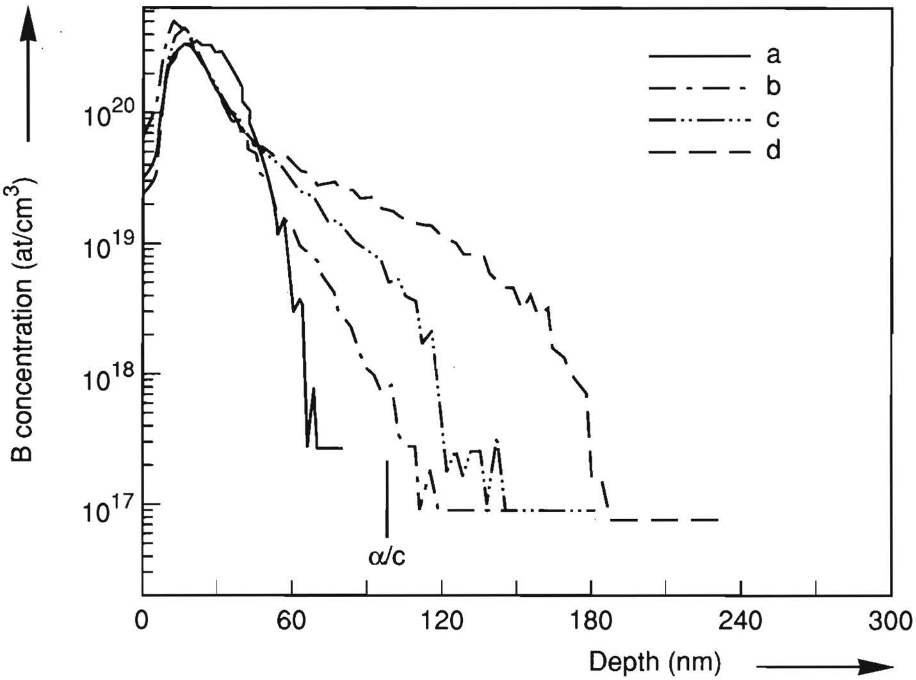


Figure 3.1. SIMS profiles of B in silicon preamorphized by Si implantation. The preamorphization conditions were Si^+ 50 keV, 1×10^{15} cm^{-2}. The B was implanted as BF_2^+ 25 keV, 1×10^{15} cm^{-2}. The annealing conditions were (a) no anneal, (b) 600 °C, 60 min plus 850 °C, 5 min, (c) 600 °C, 60 min plus 850 °C, 30 min, (d) 600 °C, 60 min plus 850 °C, 2 h. All anneals throughout this work were performed in nitrogen ambient. The position of the α/c interface is shown.

silicon as expected. Figure 3.2 shows similar results for the samples which had been preamorphized using Ge (120keV 4×10^{14} cm^{-2}) while Fig. 3.3 shows the results for BF_2^+ (25keV 1×10^{15} cm^{-2}) implanted into crystalline silicon.

There are several important points to note from Fig. 3.3. The presence of a channeling tail leading to a deeper as-implanted profile can be clearly seen when comparison is made with Fig. 3.1. There is significant diffusion of the B profile after the 600°C anneal, at concentrations below about 3×10^{17} cm^{-2}, consistent with previous observations of implantation-damage enhanced diffusion at low temperature [24]. The double hump in the peak region of the profile arises from the segregation

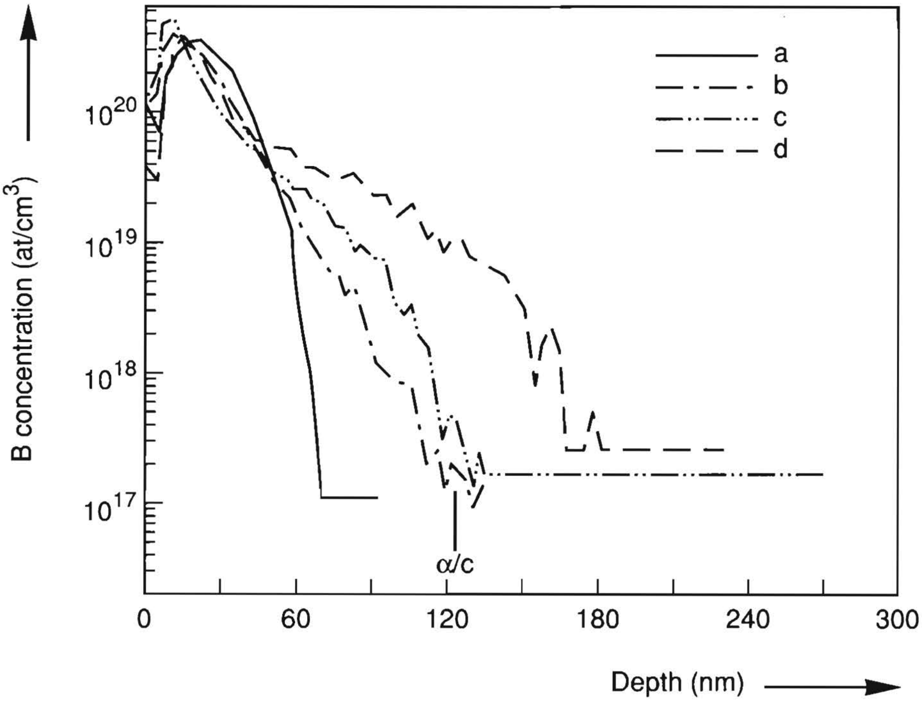


Figure 3.2. As Fig.3.1 except preamorphization carried out with Ge^+ 120 keV, 4×10^{14} cm^{-2}.

of boron to defect bands produced by the BF_2^+ implantation.

Figure 3.4 shows the result of cross-sectional transmission electron microscopy (XTEM) of the silicon surface after preamorphization by Si implantation followed by anneals at 600°C for 60 minutes and 850°C for 5 minutes. The apparent roughness and contrast variations seen in all of the XTEM micrographs are measurement artefacts due to sample thickness variations and surface roughness effects which lead to diffraction thickness fringes. The amorphous layer thickness in Fig. 3.4 was about 90 nm, which is marked in Fig. 3.1. As can be seen, the as-implanted B profile is completely contained within this amorphous layer. The thick defect band is situated at the position of the previous α/c interface. The silicon above and below this band seems to be defect-free as far as XTEM can detect. Figure 3.5 shows the equivalent microstructure after Ge implantation (120keV 4×10^{14} cm^{-2}) and annealing. The amorphous layer thickness in this case was around 125 nm, which is also marked

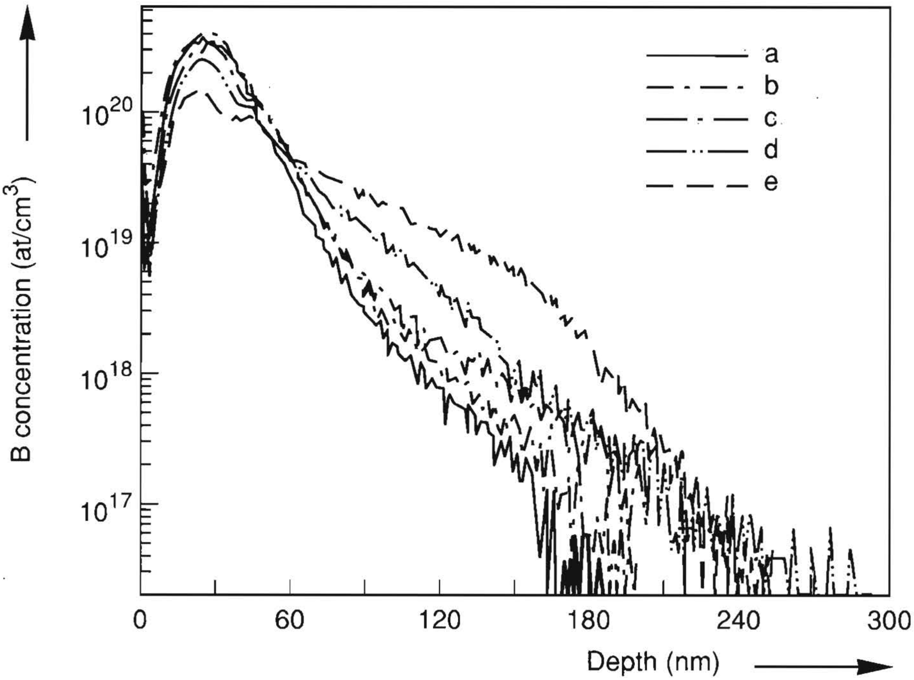


Figure 3.3. SIMS profiles of B after implantation of BF_2^+ implantation into crystal silicon. The implant conditions were BF_2^+ 25 keV, 1 × 10^{15} cm^{-2}. The anneal conditions were (a) no anneal, (b) 600 ^oC, 60 min, (c) 600 ^oC, 60 min plus 850^{o}C, 5 min, (d) 600 ^oC, 60 min plus 850^{o}C, 30 min, (e) 600 ^oC, 60 min plus 850^{o}C, 2 h.

in Fig. 3.2. Again, the as-implanted B profile is completely contained within the amorphous layer. A comparison between Figs. 3.4 and 3.5 shows that the density of residual defects is much lower in the case of the Ge implant.

The results of the electrical characterization from this first junction experiment are given in Table I along with the junction depths as determined by SIMS.

The junction depths were defined at a background doping level of 1×10^{17} cm^{-3} to avoid ambiguity due to the background noise in the SIMS signal at lower concentrations. Several trends can be noted from Table I. As expected, for a given amorphizing species (Si or Ge), the junction leakage is reduced and the ideality factor approaches unity as the junction moves further away from the defect band into the depth of

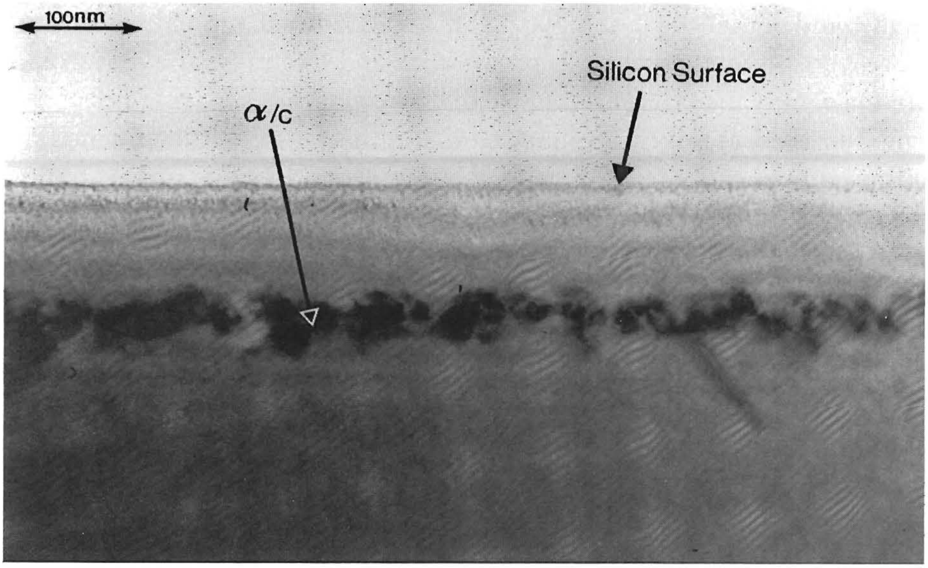


Figure 3.4. XTEM micrograph of amorphous layer after Si$^+$, 50 keV, 1×10^{15} cm^{-2} and 600 ^{o}C, 60 min plus 850 ^{o}C, 5 min.

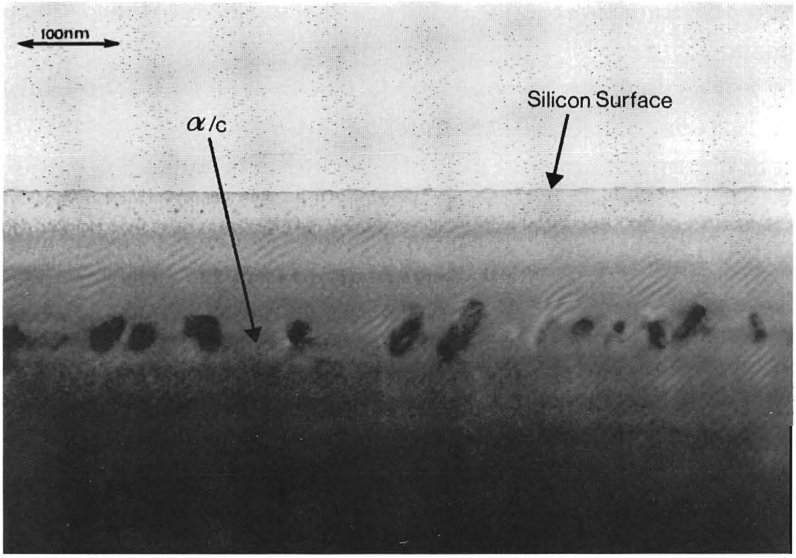


Figure 3.5. XTEM micrograph of amorphous layer after Ge$^+$, 120 keV, 4×10^{14} cm^{-2} and 600 ^{o}C, 60 min plus 850 ^{o}C, 5 min.

the silicon. Furthermore, the separation between the junction and the defect band required to achieve a low level of junction leakage seems to be greater for Si than for Ge preamorphized samples. This systematic difference may be a consequence of the relatively coarse defect band formed in the case of Si implantation, compared with that formed after Ge implantation. Note that, when the junction depth is much shallower than the preamorphization depth, the junction leakage and the ideality factor are high. This comes from the fact that the defect band at the preamorphization depth is situated in the depletion region of the junction giving rise to large generation currents.

Figures 3.4 and 3.5 also show the relatively low density of extended defects at the location of the previous α/c interface, formed in the case of Ge implantation. As has been confirmed elsewhere [23], the region beyond the α/c interface after amorphization is rich in silicon interstitials mainly due to recoil implantation of displaced host atoms. These interstitials have been shown [19] to be the cause of the growth of the extended defects seen in the XTEM micrographs after furnace annealing. With heavier ions, lower doses can be used since less damage is required to amorphize the silicon. In addition, the tail of the implantation damage profile is steeper, thus generating fewer defects beyond the amorphized layer.

As far as leakage is concerned, these defects do not have to be completely removed, but they must be situated completely within the p^+ region. Samples 3 and 11 fulfil this condition and result in ideal low-leakage junctions with a nominal depth of about 130 nm. A surprising result, shown by sample 5, is that a simple BF_2^+ implant followed by a 600°C anneal for 60 minutes results in an almost ideal junction. The nominal junction depth in this case is about 175 nm.

Diode 2 Experiment

In this case, preamorphization was performed using 70keV instead of 120keV Ge, and BF_2^+ was implanted at 20keV instead of 25 keV, to obtain even shallower p^+ junctions. The reduced BF_2^+ implantation energy also provided improved conditions for shallow junction formation using simple BF_2^+ implantation and a low-temperature thermal anneal, i.e. without preamorphization.

Figure 3.6 shows the B profiles after Ge preamorphization. Figure 3.7 shows the B profiles of the samples which had no amorphization

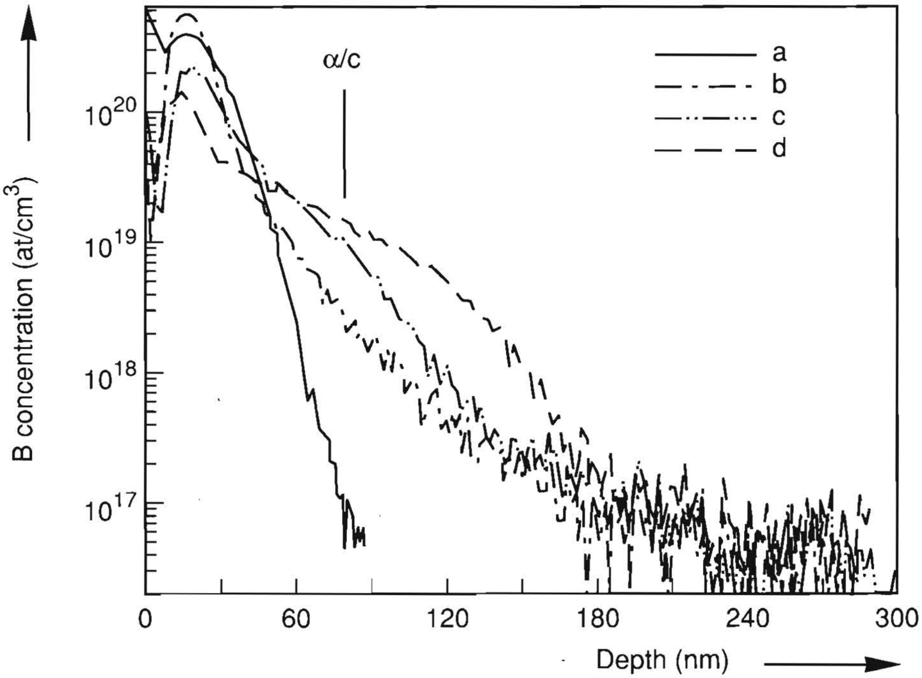


Figure 3.6. As Fig.3.1 except preamorphization carried out with Ge^+ 70 keV, $4 \times 10^{14}\ cm^{-2}$.

prior to BF_2^+ implantation.

As before, Fig. 3.6 shows four profiles while Fig. 3.7 has the added profile after 600°C annealing since this gave an ideal junction in the first experiment. Figure 3.8 shows an XTEM micrograph of the Ge preamorphized sample after annealing (600°C 60 minutes plus 850°C 5 minutes). The effect of implanting BF_2^+ at 20keV at a dose of 1×10^{15} cm^{-2} into crystal silicon can be seen in Fig. 3.9, while Fig. 3.10 shows the same sample after annealing at 600°C. The amorphous layer thicknesses in these cases have been marked in Figs. 3.6 and 3.7, being about 80 nm for the Ge implanted sample and 30 nm for the BF_2^+ implantation.

Table II gives a summary of the electrical results from this second diode experiment.

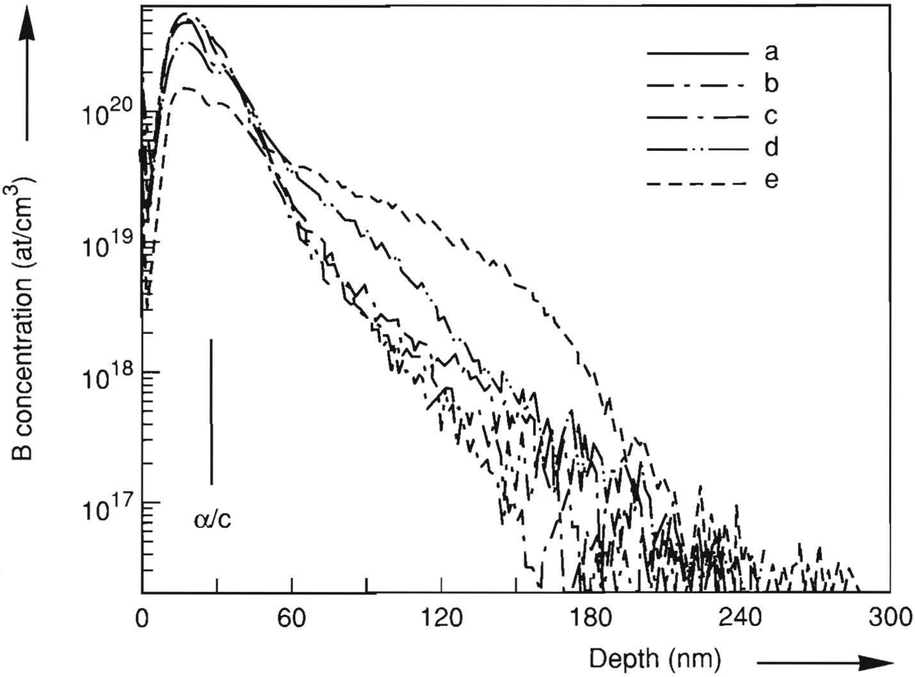


Figure 3.7. As Fig.3.3 except that implantation conditions were BF_2^+, 20 keV, 1×10^{15} cm^{-2}. The amorphous layer thickness due to BF_2^+ implantation is marked.

Sample	Ge Dose 70 keV 10^{14} cm^{-2}	Anneal (min/C) After 600C	Junc. Depth Xj (nm)	PreAmorph Depth Xa (nm)	Leak at -5V (nA/cm2)	Ideal. Fact
1	-	-	160	-	0.87	1.02
2	-	5/850	175	-	0.06	1.01
3	-	30/850	190	-	0.03	1.01
4	-	120/850	215	-	0.03	1.01
5	4	-	80	80	1.3×10^4	1.81
6	4	5/850	165	80	1.24	1.09
7	4	30/850	180	80	0.11	1.05
8	4	120/850	185	80	0.07	1.04

TABLE II : The results of the second diode experiment.

In the case of Ge preamorphization at 70keV, the tail of the as-

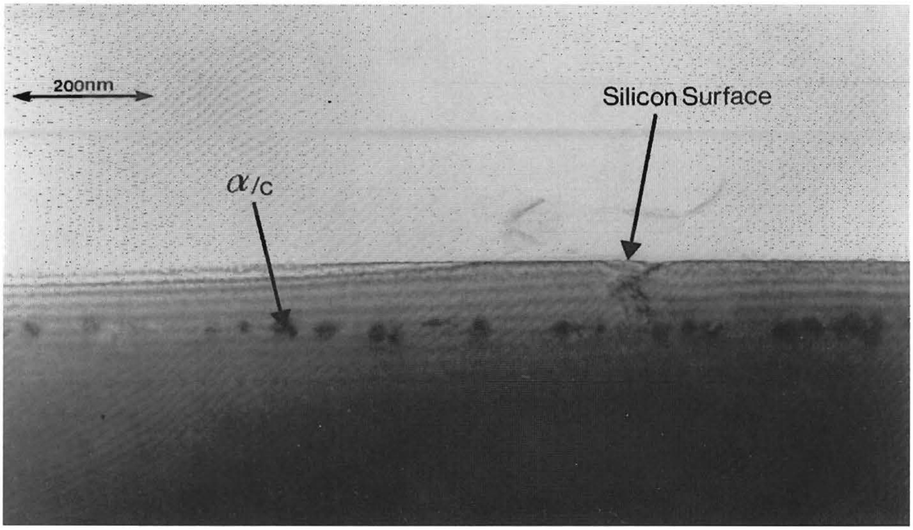


Figure 3.8. XTEM micrograph of amorphous layer after Ge$^+$, 70 keV, 4×10^{14} cm^{-2} and 600 °C, 60 min plus 850 °C, 5 min.

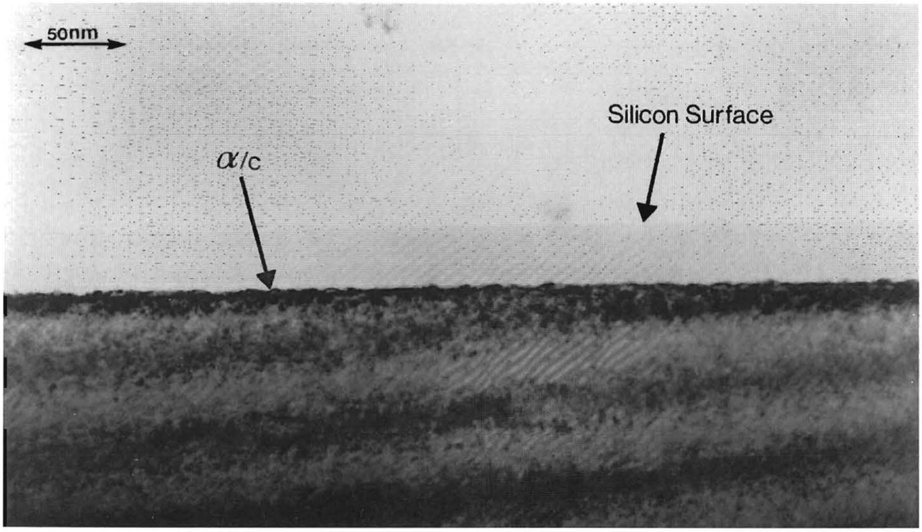


Figure 3.9. XTEM micrograph of amorphous layer after BF_2^+, 20 keV, 1×10^{15} cm^{-2}.

implanted B profile is located very close to the α/c interface. A sub-

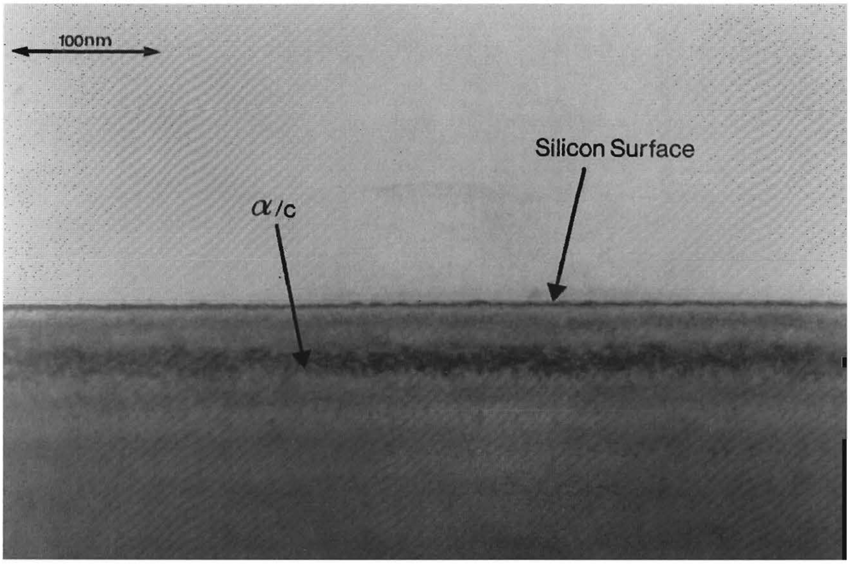


Figure 3.10. As Fig.3.9 but after 600 °C, 60 min

sequent anneal at 850°C for 5 minutes (preceded by the 600°C anneal) broadens the B profile by about 85 nm at a concentration level of 1×10^{17} cm^{-3} (samples 5 and 6, Fig. 3.6, Table II). This contrasts with a broadening of only about 40 nm for the equivalent samples in the case of Ge preamorphization at 120keV (samples 9 and 10, Fig. 3.2, Table I). Comparison of the profiles in Fig. 3.2 with those in Fig. 3.6 reveals a difference in the diffusion of the two groups. Fig. 3.6 indicates that diffusion in the tail of the B profile is initially very rapid, but slows down after the first 5 minutes at 850°C. Thereafter, the profile broadening is consistent with the normal nonlinear thermal equilibrium diffusion of B in silicon.

The situation seen in Fig. 3.6 is a consequence of transient enhanced B diffusion, caused by a supersaturation of silicon interstitials originating beyond the defect band. The smooth curvature of the B profile, with no change in slope at the defect band, suggests that in the present experiment these interstitials have been able to penetrate the defect band and cause some diffusion enhancement within the regrown layer, before recombining at bulk sinks located within this layer. In contrast, previous amorphization studies have shown no diffusion enhancement within the regrown layer [24], or even a retardation of diffusion relative to thermal equilibrium [19]). This difference suggests that the density of interstitial sinks within the regrown layer may have been significantly higher in previous studies.

Despite this difference from earlier results, the position of the as-implanted B profile within the regrown layer is clearly still critical. Although Fig. 3.2 shows significant diffusion during the first 5 minutes (almost no broadening would be expected under thermal equilibrium conditions), the observed diffusion is much less than in Fig. 3.6. This result is again consistent with the presence of sinks for silicon interstitials within the regrown layer. For a shallow junction, the tail of the as-implanted B profile has to be located at a depth significantly shallower than the original α/c interface.

The ability to make ideal low-leakage junctions using simply a BF_2^+ implantation followed by a 600^{o}C anneal is very promising and has been reported elsewhere [25]. The dopant in the tail of the B profile has evidently been activated, since otherwise the defects at the α/c interface near the peak of the profile would have contributed significant junction leakage.

3.2.3 Conclusions

Shallow p^+/n junctions have been fabricated using preamorphization and low temperature annealing. The conditions for ideal shallow junction fabrication are that the implanted B profile should be completely contained within the amorphous region and that the final junction should completely contain the defects formed at the previous amorphous-crystalline interface. Therefore, the final junction depth must be deeper than the amorphous region. This, together with the need for sharper defect bands, makes it preferable to use heavier amorphizing ion species. The limits of the technique were observed when the as-implanted B profile came too close to the α/c interface, resulting in excessive enhanced diffusion in the tail region leading to deeper junctions. The difficult nature of the process suggests it will be difficult to use as a practical fabrication method for shallow p^+ junctions.

3.3 Boron Diffusion in Preamorphized Silicon

3.3.1 Introduction

As has been emphasised before, the lateral and vertical scaling of silicon devices for very large scale integration requires the fabrication of very shallow junctions. The greatest challenge arises in the fabrication of shallow p-type junctions in n-type silicon since these p^+-n junctions require the use of boron which complicates junction formation. It is known that anomalous enhanced diffusion of boron occurs during implant damage repair which deepens the junction considerably [4], [5], [6], [7].

Nevertheless, several studies have reported the absence of anomalous enhanced diffusion in a preamorphized region [7], [26]. Guimaraes et al. [27] compared two different implanted boron doses. For a boron dose of $5 \times 10^{15}\ /cm^2$ implanted into a 200nm thick amorphous layer, enhanced diffusion was clearly seen. For the lower dose of $1 \times 10^{15}\ /cm^2$, no firm evidence for the existence of enhanced diffusion could be drawn. In the study by Kim et al. [28], boron was implanted into a 350nm thick amorphous layer. Enhanced boron diffusion was observed at 1150°C while annealing at 1000°C resulted in reduced diffusion. Xi-Mao Bao et al. [29] used a lower dose of boron, namely $1 \times 10^{14}\ /cm^2$. It was concluded that silicon interstitials coming from the damaged tail are the main driving force behind anomalous boron diffusion in the preamorphized layer. It is known from other studies [30] that the implantation of fluorine into boron implanted p^+-n junction areas leads to suppression of boron thermal redistribution. From studies on oxidation induced stacking faults in silicon [31], it is known that fault growth is repressed by fluorine due to a decrease in interstitial concentration.

In this study, the influence of rapid thermal anneal (RTA) temperature and boron dose, implanted in the form of boron difluoride (BF_2), on the diffusion of boron in preamorphized silicon is reported. Experimental results, to be presented here, have shown that boron diffusion in amorphous regions is retarded if the boron is implanted in the form of BF_2. This is in contrast to other studies where boron itself is implanted. Discussion of this effect is centred on the effect of fluorine on the boron diffusion.

3.3.2 Sample Preparation

Silicon substrates used in this experiment were $< 100 >$ n-type silicon wafers with resistivity of 10 - 15 Ω.cm. A thermal oxide of 12nm was grown on all wafers. Some wafers were implanted with silicon with a dose of 2×10^{15} $/cm^2$ and energy of 90keV. The amorphous layer thickness was about 186nm as obtained by channeled Rutherford Backscattering spectroscopy. The wafers were then implanted with BF_2 at 30keV at a dose of either 2×10^{14} $/cm^2$ or 7×10^{14} $/cm^2$. RTA was carried out at 1000^{o}C, 1100^o or 1150^{o}C for 10 seconds in nitrogen in an AG Associates 210T HEATPULSE, tungsten halogen lamp annealer. The boron profiles were measured using secondary ion mass spectrometry (SIMS) while defect structures after annealing were studied using cross sectional transmission electron microscopy (XTEM).

3.3.3 Results and Discussion

Figure 3.11a shows the boron profiles of the low dose implant (2×10^{14} $/cm^2$) into preamorphized silicon. The position of the amorphous/crystalline (α/c) interface is also given showing that the boron implant is completely contained within the amorphous layer. The apparent concentration increase at the surface is an artefact of SIMS and is due to the increased secondary ion yield due to the presence of thermal oxide. This is present in all subsequent SIMS profiles. The startling result shown in Fig. 3.11a indicates there is practically no thermal diffusion of boron for this implantation dose in preamorphized silicon. Figure 3.11b shows the same implant and anneals but carried out in crystalline silicon. In this case, the as-implanted boron profile shows channeling as expected. Subsequent annealing leads to a discernible saturation effect in the tail diffusion, indicative of transient enhanced diffusion, along with boron diffusion at higher concentrations. Figure 3.12a shows the effects of RTA's on the higher dose BF_2 implant (7×10^{14} $/cm^2$) in preamorphized silicon. In this case, thermal diffusion occurs regularly at all concentrations. Figure 3.12b shows the equivalent profiles in crystalline silicon. In this case, the saturation effect in the tail diffusion is clear, indicating enhanced diffusion.

XTEM micrographs of the preamorphized samples after 1000^{o}C and 1150^{o}C are shown in Fig. 3.13a and Fig. 3.13b respectively. Both had also been implanted with BF_2 at the lower dose prior to annealing.

The concentration of defects near the previous α/c interface certainly decreases as a result of annealing at the higher temperature. This is consistent with results obtained in [28] where it was postulated that silicon interstitials are emitted from this region when annealing is done at 1150°C. However, in this case, this has apparently no effect on the boron diffusion as seen in Fig. 3.11a. The apparent roughness and contrast variations seen in Fig. 3.13b are purely measurement artefacts due to the sample being wedge shaped which leads to diffraction thickness fringes and sample surface roughness effects.

Figure 3.14a shows the fluorine profiles in the case of the higher dose BF_2 implant in the preamorphized case. The fluorine peak at a depth of about 190nm for an RTA of 1000°C coincides with the damage at the previous α/c interface. As the RTA temperature is increased, the amount of fluorine at this depth decreases. The fluorine concentration at the implant peak also decreases with increasing temperature. The apparent increase in fluorine concentration near the surface after annealing is not an artefact of the SIMS measurement because it is not present in the as-implanted profile. These results are consistent with other publications [32] where it was reported that most fluorine atoms diffuse out of the silicon substrate with the rest remaining at residual defect sites. It is proposed that the fluorine, which has been shown to collect at the previous α/c interface during the 1000°C anneal, is emitted during higher temperature anneals and diffuses towards the surface. In this way, it passes through the tail region of the boron profile. This contrasts with the crystalline case, shown in Fig. 3.14b. In this case, all RTA anneals lead to the fluorine either diffusing out of the silicon substrate or collecting at the previous α/c interface created by the BF_2 implant itself.

As can be seen, the fluorine concentration in the tail of the boron profile reduces to negligible amounts even after the lowest temperature anneal of 1000°C. This contrasts with Fig. 3.14a which shows that the fluorine concentration in the tail of the boron profile decreases more gradually as the temperature increases due to fluorine emission from the "reservoir" at the previous α/c interface. Figure 3.15 shows the SIMS fluorine profiles in the case of the low dose implantation into preamorphized silicon. In this case, the as-implanted, 1000°C and 1100°C samples are shown. The same behaviour is seen as in Fig. 3.14a for the higher dose sample. The same fluorine reservoir forms at the defects at the previous amorphous/crystalline interface after the lowest temperature anneal. Higher temperatures result in fluorine emission from this region.

3.3.4 Conclusions and Summary

The complete absence, shown here, of enhanced diffusion of boron in preamorphized silicon at temperatures up to 1150°C is proposed to be due to the effect of fluorine on the boron diffusion characteristics. The presence of a fluorine "reservoir" at the α/c interface which emits fluorine towards the silicon surface clearly inhibits the boron diffusion as long as the boron profile is contained within the amorphous region. As has been seen elsewhere [31], the presence of fluorine in silicon results in a decrease in the interstitial concentration. It is envisaged that the emission of fluorine from the reservoir situated at the previous amorphous/crystalline interface results in a decrease in the interstitial concentration in the boron profile tail. This, in turn, will lead to a reduced diffusion of boron as has been shown here.

In summary, it has been demonstrated that the diffusion of boron in preamorphized silicon is substantially reduced with the complete absence of anomolous enhanced diffusion if the boron is implanted in the form of boron difluoride. It was proposed that this is due to the effect of fluorine on the interstitial concentration.

3.4 General Conclusions

The ability to fabricate shallow junctions in silicon is of crucial importance for the continued miniaturisation of devices. The PMOS device, in particular, presents difficulties due to the necessity to use boron as the dopant for source and drain. The ability to make ideal shallow p^+ junctions using preamorphization and a low temperature thermal cycle was shown indicating that the inevitable crystal defects arising from the technique do not have to be completely removed. With the unavoidable reduction in thermal cycles, the presence of these defects would have to be taken into account in future manufacturing if this technique were to be employed. It was stated, however, that due to its difficult nature, preamorphization will probably be a difficult technique to apply in a production environment. In the case of junctions shallower

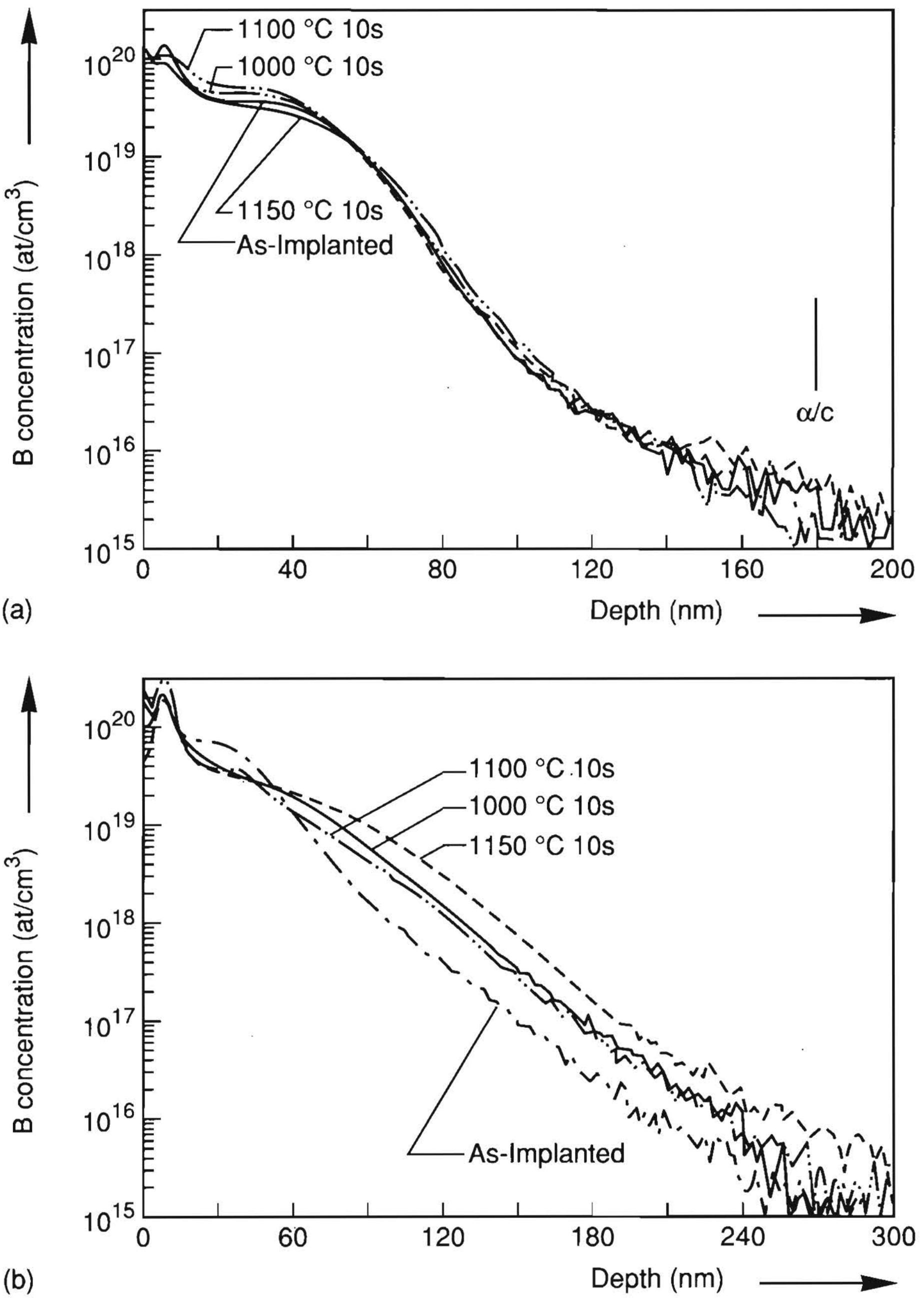


Figure 3.11. SIMS boron profiles in (a) preamorphized and (b) crystalline samples implanted with 30 keV BF_2^+ at a dose of 2×10^{14} cm^{-2} after RTA.

than 0.1μm, special techniques will have to be used. These include

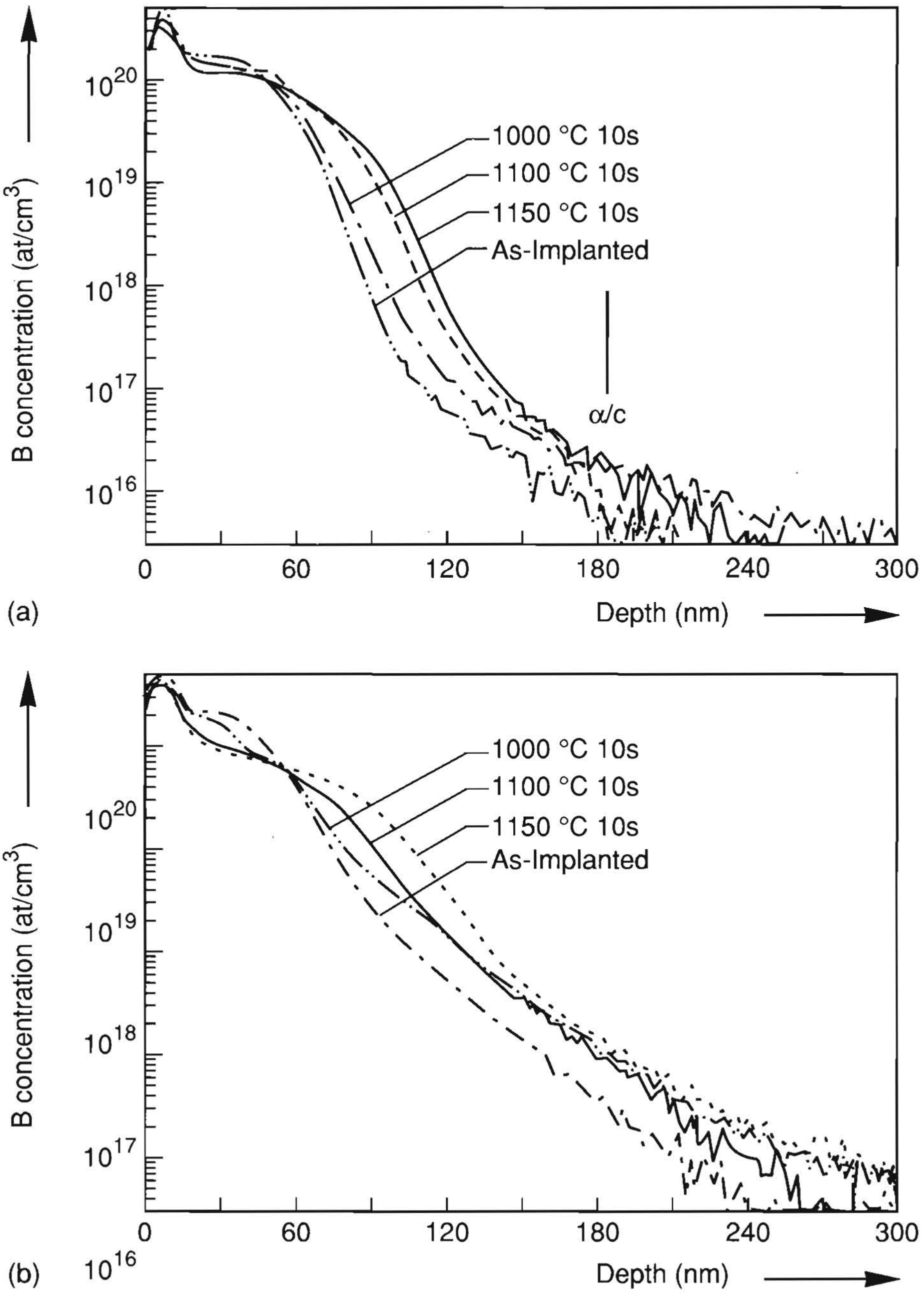


Figure 3.12. SIMS boron profiles in (a) preamorphized and (b) crystalline samples implanted with 30 keV BF_2^+ at a dose of $7 \times 10^{14}\ cm^{-2}$ after RTA.

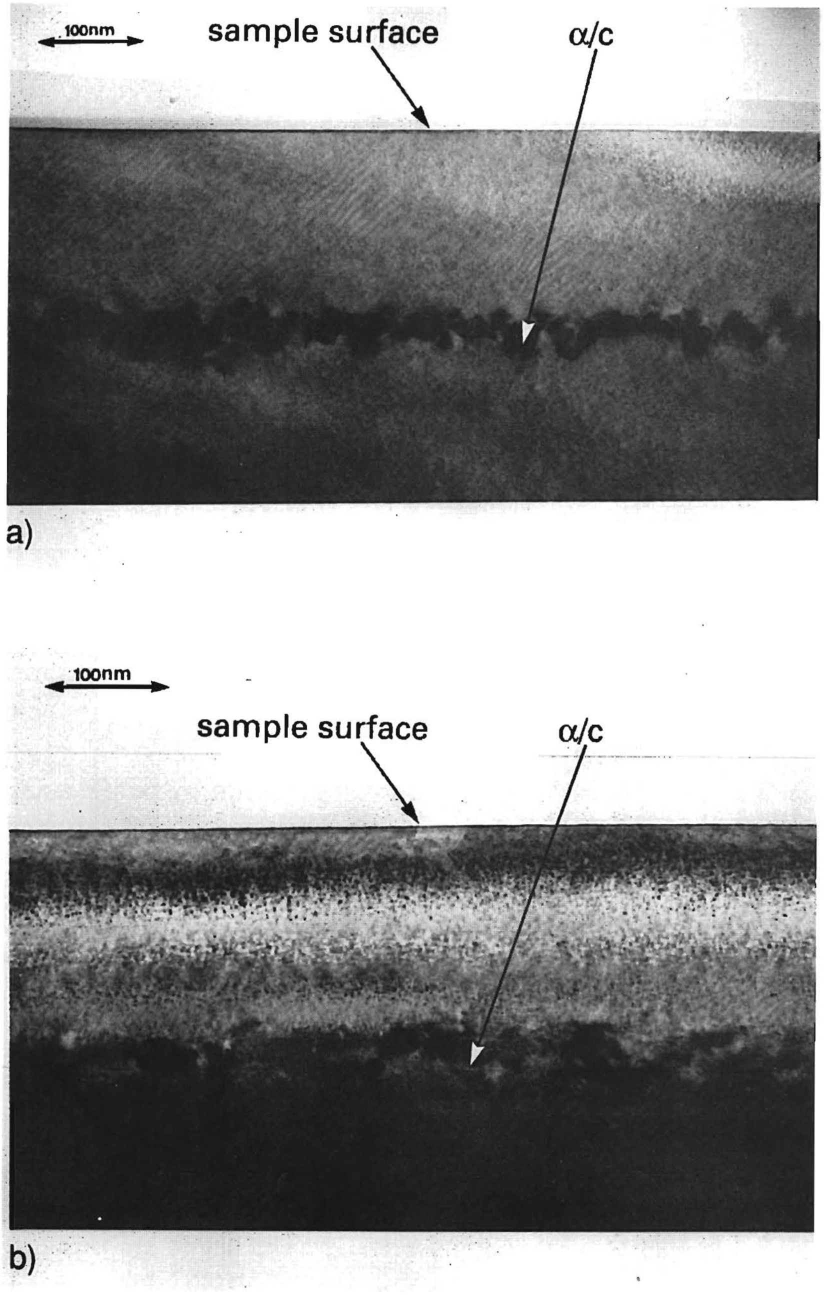


Figure 3.13. XTEM micrographs of preamorphized samples (Si^+, 2×10^{15} cm^{-2}, 90 keV) implanted with BF_2^+, 30 keV, 2×10^{14} cm^{-2} after (a) 1000 ^{o}C, 10 s and (b) 1150 ^{o}C, 10 s. The positions of the silicon surface and the previous amorphous/crystalline (α/c) interface are shown.

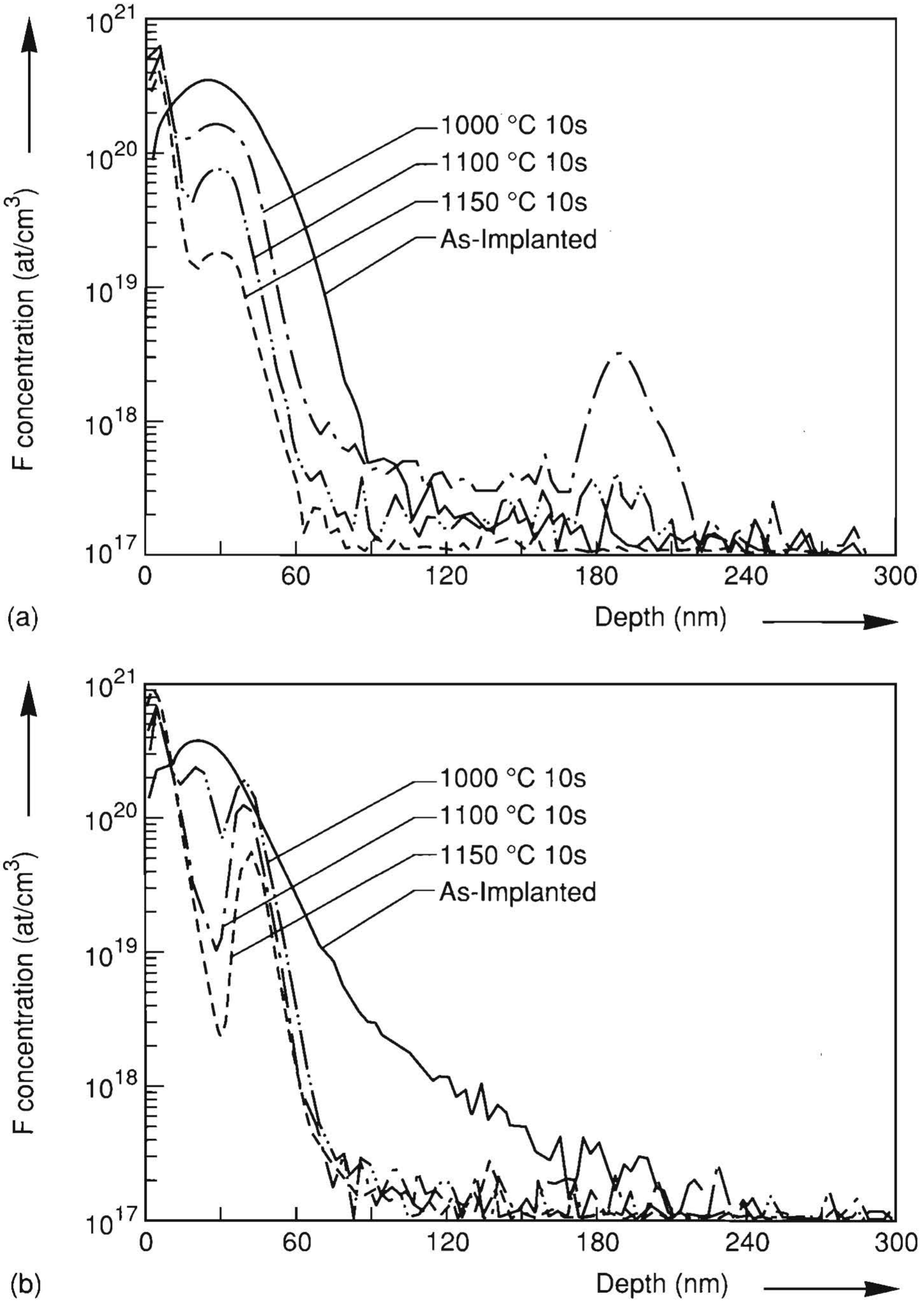


Figure 3.14. SIMS fluorine profiles in (a) preamorphized and (b) crystalline samples implanted with 30 keV BF_2^+ at a dose of 7×10^{14} cm^{-2} after RTA.

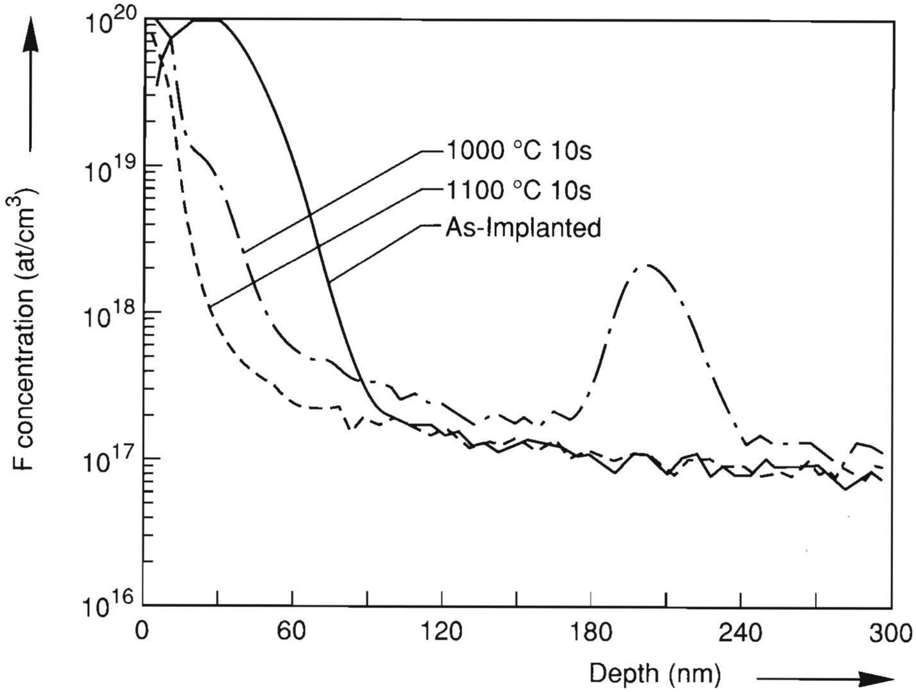


Figure 3.15. SIMS fluorine profiles in preamorphized sample implanted with 30 keV BF_2^+ at a dose of 2×10^{14} cm^{-2} after RTA.

diffusion from solid sources such as amorphous and polycrystalline silicon [33] leading to elevated source and drain MOSFETs [34]; rapid thermal diffusion from a spin-on source [35]; diffusion from silicides [36]; low energy ion implantation and rapid thermal annealing [37]. As always, manufacturability will eventually define which technique is ultimately used for mass production.

Results were presented of boron diffusion in the presence of fluorine in the amorphous region formed by heavy ion implantation. The fluorine was implanted along with the boron in the form of boron difluoride. It was seen that the fluorine congregates at the defect band formed at the previous amorphous - crystalline interface. It was proposed that this acts as a reservoir of fluorine which is released at high temperatures. This proposal would explain the startling result of almost complete absence of boron diffusion at temperatures up to 1150°C and at low boron fluences.

The study of shallow boron junction fabrication presents a challenging area which will define whether further miniaturisation is possible in silicon technology.

3.5 References

[1] T.M. Liu and W.G. Oldman, IEEE Electron Device Lett., vol. EDL-4, p 59 1983.

[2] A.E. Michel, R.H. Kastl, S.R. Mader, B.J. Masters and J.A. Gardner, Appl. Phys. Lett. 44 (4), pp.404-406, Feb. 1984.

[3] S.S. Cohen, J.F. Norton, E.F. Koch and G.J. Weisel, J. Appl. Phys. 57 (4), pp.1200-1213, Feb. 1985.

[4] R.T. Hodgson, V.R. Deline, S. Mader and J.C. Gelpey, Appl. Phys. Lett. 44 (6), pp.589-591, March 1984.

[5] K. Cho, M. Numan, T.G. Finstad, W.K. Chu, J. Liu and J.J. Wortman, Appl. Phys. Lett. 47 (12), pp.1321-1323, Dec. 1985.

[6] M. Servidori, R. Angelucci, F. Cembali, P. Negrini and S. Solmi, J. Appl. Phys. 61 (5) pp. 1834-1840, March 1987.

[7] T.O. Sedgwick, A.E. Michel, V.R. Deline, S.A. Cohen and J.B. Lasky, J. Appl. Phys. 63 (5) pp. 1452-1463, March 1987.

[8] A.J. Walker, J. Appl. Phys. 71 (4) pp. 2033-2035, February 1992.

[9] A.J. Walker, P.H. Woerlee, H.G. Pomp and N.E.B. Cowern, J. Appl. Phys. 73 (8) pp. 4048-4053, April 1993.

[10] M.Y. Tsai and B.G. Streetman, J. Appl. Phys., 50 , pp. 183-187, Jan. 1979.

[11] B-Y. Tsaur and C.H. Anderson, Jr., J. Appl. Phys. 54 (11), pp. 6336-6339, Nov. 1983.

[12] M. Delfino, D.K. Sadana and A.E. Morgan, Appl. Phys. Lett., 49 (10), pp. 575-577, Sept. 1986.

[13] I.D. Calder, H.M. Naguib, D. Houghton and F.R. Shepherd, Mat. Res. Soc. Symp. Proc. Vol.35, 1985.

[14] L. Mader, M. Orlowski and I. Weitzel, Proc. of the European Solid State Dev. Res. Conf., Berlin, Germany, pp. 41-44, Sept. 1989.

[15] P. Mei, B. Jalali, E.S. Yang, N.G. Stoffel and D.L. Hart, Appl. Phys. Lett. 56 (14), pp. 1362-1364, April 1990

[16] E. Ganin, G.A. Sai-Halasz and T.O. Sedgwick, Mat. Res. Soc. Symp. Proc. Vol.92, pp. 33-38, 1987.

[17] A.C. Ajmera and G.A. Rozgonyi, Appl. Phys. Lett., vol.49, pp. 1269-1271, Nov. 1986.

[18] M.C. Ozturk, J.L. Wortman, C.M. Osburn, A. Ajmera, G.A. Rozgonyi, E. Frey. W-K. Chu and C. Lee, IEEE Trans. Electron Dev. Vol.35, pp. 659-668, May 1988.

[19] S.D. Brotherton, J.R. Ayres, J.B. Clegg and J.P. Gowers, J. of Electronic Materials, Vol.18, pp. 173-184, 1989.

[20] A. Tanaka, T. Yamaji, A. Uchiyama, T. Hayashi, T. Iwabuchi and S. Nishikawa, IEDM Tech. Dig., pp. 785-788, 1989.

[21] E. Ganin, B. Davari, D. Harame, G. Scilla and G.A. Sai-Halasz, Appl. Phys. Lett. 54 (21), pp. 2127-2129, May 1989.

[22] F. Cembali, M. Servidori, E. Landi and S. Solmi, Phys. Status Solidi (a), vol.94, pp. 315-319, 1986.

[23] J. Thornton, K.C. Paus, R.P. Webb, I.H. Wilson and G.R. Booker, J. Phys. D: Appl. Phys. 21, pp.334-338, 1988.

[24] N.E.B. Cowern, K.T.F. Janssen and H.F.F. Jos, J. Appl. Phys. 68, 6191 (1990).

[25] S. Tatsuta and T. Sakurai, Fujitsu Scientific and Technical Journal, 14(4), pp.73-82, 1978.

[26] M.C. Ozturk and J.J. Wortman, Appl. Phys. Lett. 53, p 281, 1988.

[27] S. Guimaraes, E. Landi and S. Solmi, Phys. Stat. Sol. (a) 95, pp.589-598, 1986.

[28] Y.M. Kim, G.Q. Lo, D.L. Wong, H.H. Tseng and R. Hance, Appl. Phys. Lett. 55 (22), pp. 2316-2318, Nov. 1989.

[29] Xi-Mao Bao and Xue-Mei Hua, Phys. Stat. Sol. (a) 123, K89, Feb. 1991.

[30] K. Ohyu, T. Itoga and N. Natsuaki, Jap. J. of Appl. Phys., 29, pp. 457-462, March 1990.

[31] F. Kuper, Ph.D thesis, University of Groningen, The Netherlands, 1987.

[32] R.G. Wilson, J. Appl. Phys. 54 p6879 1983

[33] K. Park, S. Batra, S. Banerjee and G. Lux, J. Electronic Materials, vol. 20, pp. 261-265, 1991.

[34] S.S. Wong, D.R. Bradbury, D.C. Chen and K.Y. Chiu, IEDM Tech. Dig. pp. 634-637 1984.

[35] A. Usami, M. Ando, M. Tsunekane and T. Wada, IEEE Trans. Elec. Dev., ED-39, pp. 105-110, 1992.

[36] H. Jiang, C.M. Osburn, Z.G. Xiao, G. McGuire, G.A. Rozgonyi, B. Patnaik, N. Parikh and M. Swanson, J. Electrochemical Soc., vol. 139, pp. 206-211, 1992

[37] S.N. Hong, G.A. Ruggles, J.J. Wortman and M.C. Ozturk, IEEE Trans. Elec. Dev., ED-38, pp. 476-486, 1991.

Chapter 4

Gate Work Function Engineering

"Life is only a very dull and ill-directed theatre unless we have some interests in the piece; and to those who have neither art nor science, the world is a mere arrangement of colours, or a rough footway where they may very well break their shins."
R.L. Stevenson

4.1 Introduction

As MOS transistor gate lengths approach $0.5\mu m$, the work function of the gate electrode becomes increasingly important. In the standard procedure for complementary MOS (CMOS), the gate electrode is formed from polysilicon degenerately doped with phosphorus. To obtain the required threshold voltages for the PMOS and NMOS devices, a p-type implant is required in the channel regions of both types of devices. This results in a "surface channel" NMOS device where the threshold voltage implant is of the same type as the substrate, and a "buried channel" PMOS device where the implant is of the opposite type to the substrate. In this latter case, a channel p-n junction is formed. As has been shown elsewhere [1], this type of device is more susceptible to short-channel effects such as threshold voltage roll-off and drain-induced barrier lowering [2]. As devices shrink even more, the supply voltage must be lowered for reliability and operational considerations [3]. This means that the threshold voltages of the transistors must be reduced in the same manner to conserve current drive capabilities.

The consequence for the PMOS device is that heavier compensating implants will be required in the channel region leading to deeper channel p-n junctions and more short-channel problems. Various methods have been attempted to avoid these effects. These include retrograde N-well formation [4], threshold voltage increase [5], and the formation of very shallow p^+-n junctions [6].

An alternative approach is to change the work function of the gate material itself. To appreciate the effect this has on the threshold voltage of a PMOS device, the following calculation is shown [7].

Assuming a uniformly doped substrate, the PMOS threshold voltage is given as

$$V_T = \Phi_{MS} - Q_o/C_{ox} + \Phi_B - Q_B/C_{ox} \tag{4.1}$$

where Φ_{MS} is the work function difference between the gate electrode and the substrate, Q_o is the oxide charge per unit surface area, C_{ox} is the gate oxide capacitance per unit surface area, Φ_B is the surface potential of the substrate, and Q_B is the depletion charge per unit surface area when the surface has reached inversion.

Looking at the two possible cases:

- **N-type polysilicon**

  In this case, Φ_{MS} = -0.28 V, Φ_B = -0.75 V. The contribution from the fixed oxide charge can be neglected since a value of less than 10mV is contributed by the Q_o/C_{ox} term when Q_o is taken as $10^{10}e$ C/cm^2 (e = electronic charge), a value typical in modern process technology, and the gate oxide thickness is 15 nm. Therefore,

  $$V_T = -1V - Q_B/C_{ox}$$

  Therefore, to obtain a V_T of about -0.6 V, Q_B must be made negative. In other words, the surface must be p-type which is opposite to the dopant in the substrate.

- **P-type polysilicon**

In this case, $\Phi_{MS} = 0.84$ V, $\Phi_B =$ -0.75 V. Again, the contribution from fixed oxide charge can be neglected. Therefore,

$$V_T = 0.1V - Q_B/C_{ox}$$

Thus, a positive Q_B would be necessary to obtain a V_T of about -0.6 V. That is, an n-type implant must be carried out in the channel region which is then of the same dopant type as the substrate.

From the above, it can be seen that a PMOS device with a p-type polysilicon gate electrode is the mirror image, as it were, of the NMOS device with an n-type gate. That is, the two devices are completely symmetrical. Besides avoiding the channel p-n junction in the PMOS device, the use of p-type polysilicon as the gate electrode requires using an n-type channel implant such as arsenic which has a low thermal diffusivity in silicon compared to boron. Therefore, as far as the channel profile is concerned, the surface channel device can withstand higher processing temperatures compared to the buried channel device. Unfortunately, this is not the whole story since previous studies have shown increasing boron penetration from the gate to the channel region with increasing anneal temperatures [8] and annealing with hydrogen present [9]. To fabricate a p-type gate, boron is used and can be introduced in two ways; it can be implanted as ionic boron, B^+ or as boron difluoride, BF_2^+. Further studies have shown that fluorine plays a crucial role in boron penetration from the gate into the PMOS channel region [10], [11].

The work reported here looks at the successful and reliable use of B^+ doped polysilicon for the gates of submicron PMOS devices [12]. The maximum thermal budget for negligible boron penetration from the gate electrode was investigated along with a positive bias temperature stress instability in which gate oxide interface states increase in number when a positive bias is applied to the gate with respect to the substrate. Both n-type and p-type polysilicon gated devices were fabricated to enable a direct comparison to be made between the two concerning current-voltage characteristics and lifetime predictions.

4.2 Experimental

Two FZ $< 100 >$ n-type 10-15 Ω.cm batches of wafers were processed for this study. The first involved the fabrication of large capacitor structures to study boron penetration of the gate oxide and gate oxide stability. In this, polysilicon gate electrodes were formed by deposition and etching upon a 12.5 nm gate oxide. Doping of the polysilicon was done by implantation as follows: (I) some wafers were implanted with phosphorus ($5 \times 10^{15} cm^{-2}$) and annealed at 900°C in dry nitrogen; (II) others were similarly implanted with 16keV boron ($5 \times 10^{15} cm^{-2}$) and annealed; (III) some of the wafers from (II) were further implanted with phosphorus ($10^{16} cm^{-2}$) and annealed. Therefore, there were three types of gate electrodes:

I n-type with only phosphorus present,

II p-type with only boron present and, finally,

III n-type with both phosphorus and boron.

In (III), the boron is distributed in the polysilicon and gate oxide in the same way as in (II) where there is only boron present. The only difference is that in (III), the polysilicon is overdoped with phosphorus to make the gate n-type. For some wafers, where the polysilicon had been doped only with phosphorus or boron, different anneals were carried out after polysilicon doping to test boron penetration. These were 850°C, 900°C, 925°C, 950°C all in dry nitrogen.

The second batch involved the fabrication of PMOS devices with minimum polysilicon gatelengths of $0.7\mu m$ and minimum effective channel lengths of $0.4\mu m$. The surface channel device had surface and well concentrations of $2 \times 10^{17} cm^{-3}$ and $5 \times 10^{16} cm^{-3}$ respectively, while the buried channel device had an arsenic anti-punchthrough concentration of $2 \times 10^{17} cm^{-3}$. A 12.5 nm gate oxide was grown followed by a threshold voltage adjust implant. This took the form of a boron difuoride implant for the buried channel devices and an arsenic implant for the surface devices. Polysilicon was then deposited and doped in the case of the n-type polysilicon wafers with a phosphorus implant ($10^{16} cm^{-2}$). After patterning and etching, the source/drain implant (boron $5 \times 10^{15} cm^{-2}$ 16keV) was carried out doping the polysilicon in the case of the p-type polysilicon wafers. Annealing, dielectric deposition and metalisation were followed by a forming gas anneal at 450°C.

The transistors were stressed at maximum gate current (-1.5V source/gate voltage) with various source/drain voltages. with the maximum linear transconductance, g_m, being monitored since this parameter changed most rapidly. The device was defined to be out of specification if g_m had changed (increased) by ten percent. Extrapolation of the lifetime versus inverse source/drain voltage to 30 years then gives the maximum V_{ds} for 30 years operation within specification.

4.3 Results and Discussion

Figure 4.1 shows the difference in flatband voltage for the capacitor samples (I) and (II), measured using the high frequency capacitance - voltage method, as a function of thermal budget after implantation doping of the polysilicon. 900°C 60 minutes in dry nitrogen seems to be the maximum budget to limit boron diffusion through the 12.5nm gate oxide.

The effect of a positive bias temperature stress (+5V, 250°C, 5min) can be seen in Fig. 4.2 in which quasi-static C-V curves are shown from capacitors with n- and p-type gates. Only the type (II) capacitors with p-type gates were affected and only with a positive voltage applied to the gate electrode. This was true even when boron was present in the gate oxide and the polysilicon as in (III). Therefore, this instability seems to be independent of whether boron is present in the polysilicon or gate oxide. Rather, it appears that the polysilicon work function is the important parameter determining the gate oxide stability.

The reasons for this can be speculated upon. In [13], it was stated that the only mechanism possible to explain the positive bias instability is hole injection from the p^+ gate during the stress. These subsequently travel to the Si/SiO_2 interface and generate interface states. How this is done was not given. As was discussed in [14], it is questionable whether there is any hole transport in the gate oxide. It was speculated that photon production is an important mechanism. This would occur in the following manner. Electrons are injected from the cathode, which is the substrate in the case under discussion, and travel in the conduction band of the gate oxide. On reaching the gate electrode, they drop from the oxide conduction band to empty states near the gate electrode's

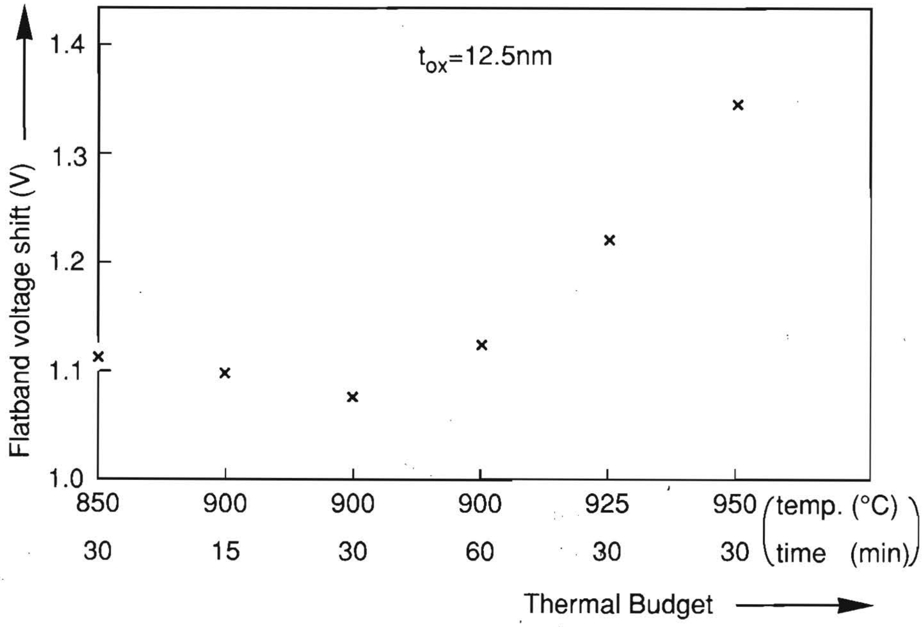


Figure 4.1. Flatband voltage as a function of thermal budget applied after polysilicon doping. The flatband voltage was measured from the high frequency capacitance-voltage curve.

Fermi level. The potential energy of the electron in the oxide conduction band is then converted to some other form. This could be the breaking of bonds at the gate/oxide interface and/or the production of a photon. The photon, which is hardly absorbed by the gate oxide, can reach the other interface and cause bond breaking and interface state generation. It thus becomes clear why the situation is so much worse in the case of a p^+ gate. Electrons falling from the conduction band of the oxide into the p^+ gate will have to lose 1.1 eV more energy than those falling into the n^+ gate.

Figure 4.3 shows the increase of interface states for a p-poly capacitor as a function of inverse absolute stress temperature. The activation energy of 0.9eV agrees quite well with [13].

Figure 4.4 shows the linear and saturation currents in the transistors with either n- or p-type polysilicon gates. The 15% lower current drive capability of the p-poly pMOS can be seen and is due to the lower hole mobility caused by the higher effective field in the inversion layer

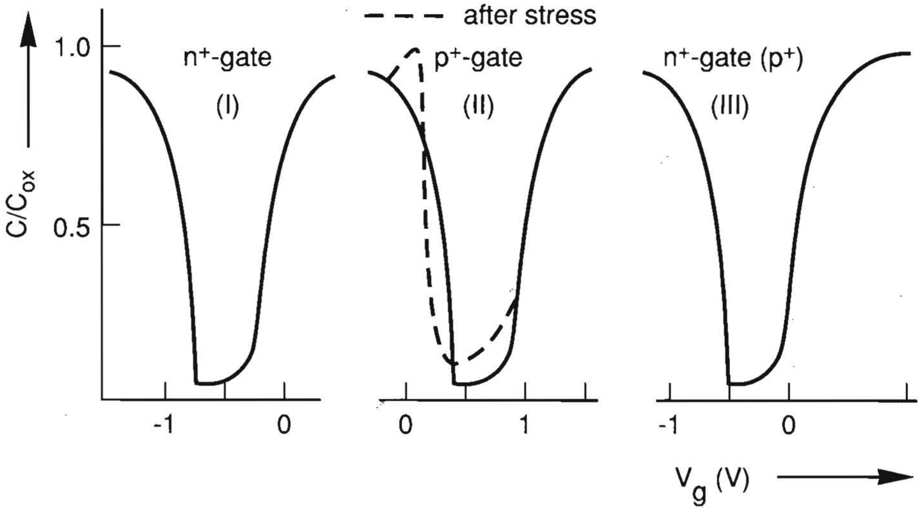


Figure 4.2. Quasistatic capacitance-voltage curves for the three types of capacitor structures before and after a positive bias temperature stress. The three types of capacitor are (I) n-type with only phosphorus present, (II) p-type with only boron present and, finally, (III) n-type with both phosphorus and boron.

[15]. The transistors were stressed as described above.

Figure 4.5 shows the transistor lifetime (time required for a 10% shift in the linear transconductance). Extrapolation to 30 years operation gives the maximum source/drain voltage for 30 years operation within specification.

Figure 4.6 shows this voltage plotted as a function of effective channel length for both types of pMOS transistor. As is evident, the surface channel transistors can withstand at least 0.5 volt more than the buried channel devices. Similar results have been reported elsewhere [16], [17].

Such a result can be explained by looking at the maximum gate current as a function of source-drain voltage, V_{ds}, for the two types of devices as shown in Fig. 4.5.

The larger gate current in the buried channel case is not only due to the larger source-drain current but is a result of the required large antipunchthrough implant which, by increasing the fields, increases the gate and substrate currents [18]. This effect becomes even worse as buried channel devices are further reduced in size due to the heavier

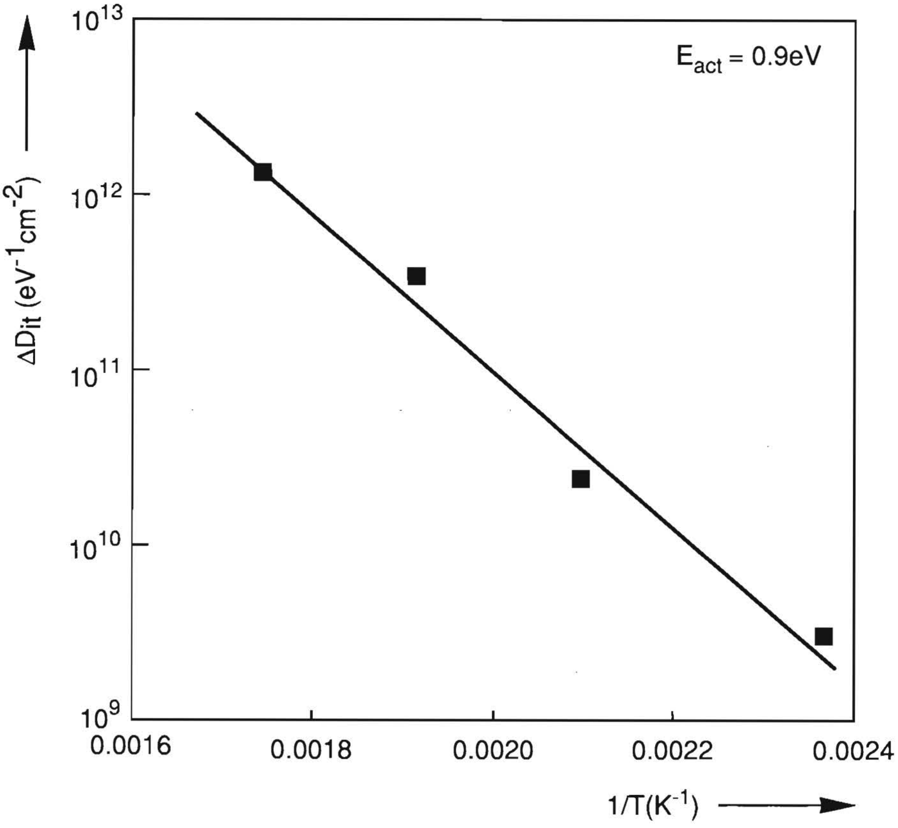


Figure 4.3. Interface state increase as a function of inverse absolute temperature at which the stress took place.

implants required. In addition, the larger work function of the p-poly gate electrode increases the normal field in the oxide compared to the n-poly case pulling more electrons towards the gate electrode.

4.4 Conclusions

As buried channel PMOS devices are miniaturized, they become more susceptible to short channel and hot carrier degradation effects. Despite performance reduction and interface state density increase with positive bias, it has been shown that the use of p-type polysilicon can be used to fabricate submicron PMOS devices which are more resistant to hot carrier degradation. It is anticipated that such surface channel

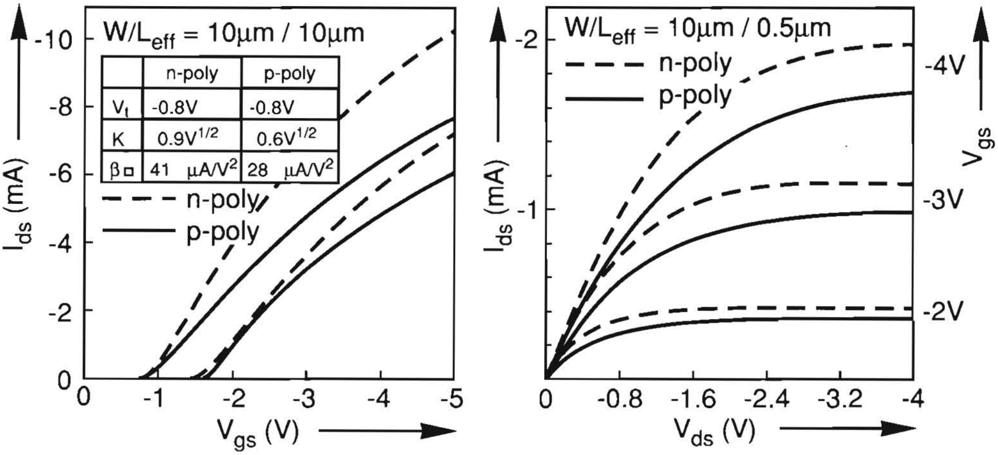
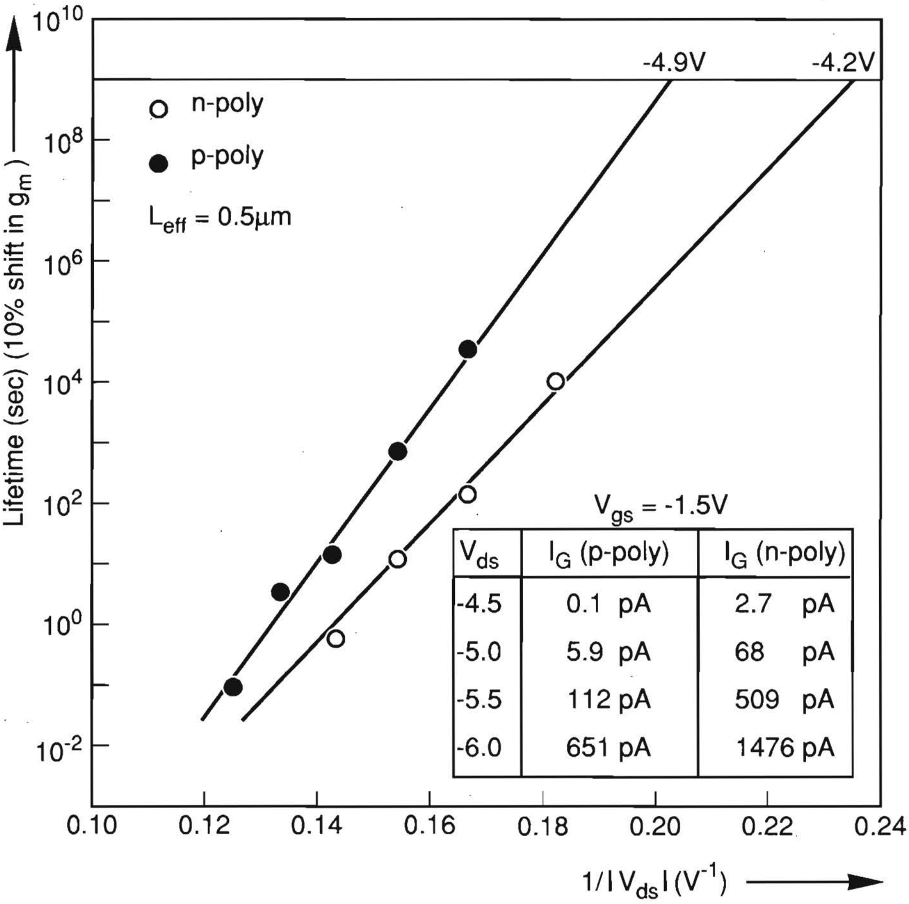


Figure 4.4. Linear and saturation current-voltage curves for the transistors with either n- or p-type polysilicon gates.

PMOS devices will become indispensable as device gate lengths are reduced in size to below 0.5μm. Future work should concentrate on the use of boron difluoride (BF_2^+) implantation to simultaneously dope the PMOS gate polysilicon and form the source and drain. As was seen in chapter 3, BF_2^+ implantation followed by low temperature annealing resulted in ideal junctions. As has been stated, more recent publications have shown that the use of BF_2^+ to dope the gate polysilicon is a more critical process than when using boron alone [11], [19]. This has been attributed to the presence of fluorine which enhances boron diffusion through the gate oxide. More recent work has looked at the use of BCl implantation since chlorine does not seem to enhance the boron diffusion through the gate oxide [20]. The most promising approach seems to be the use of nitrided thin oxides as gate dielectric [21] since the nitridation seems to block boron penetration towards the transistor channel region.



V_{ds}	I_G (p-poly)	I_G (n-poly)
-4.5	0.1 pA	2.7 pA
-5.0	5.9 pA	68 pA
-5.5	112 pA	509 pA
-6.0	651 pA	1476 pA

Figure 4.5. The transistor lifetime (time required for a 10% shift in the linear transconductance). Extrapolation to 30 years operation gives the maximum source/drain voltage for 30 years operation within specification.

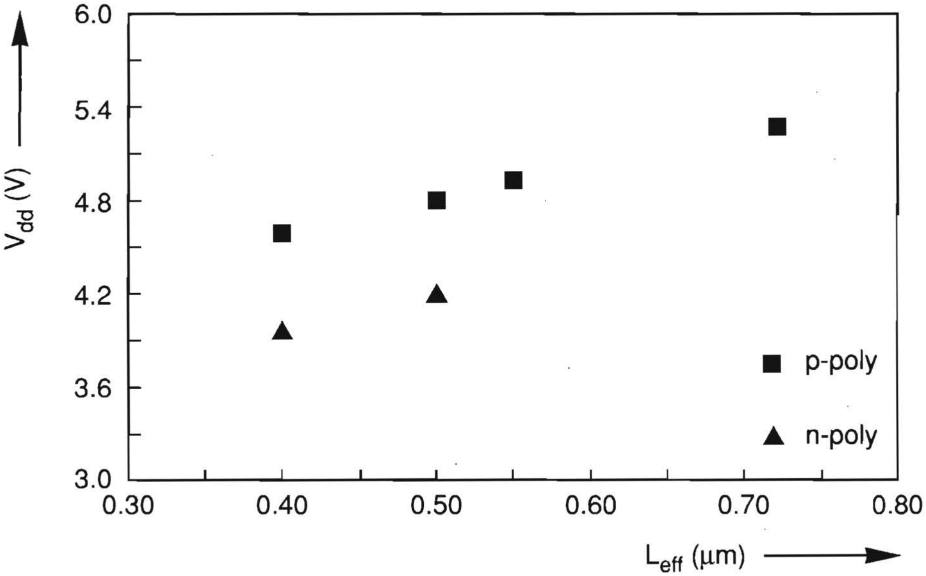


Figure 4.6. Maximum source/drain voltage for 30 years operation within specification, as a function of effective channel length.

4.5 References

[1] M.R. Wordeman, IEDM Tech. Dig., pp.26-29, Dec. 1979.

[2] G.J. Hu and R.H. Bruce, IEEE Trans. Elec. Dev., ED-32, pp. 584-588, March 1985.

[3] see, for instance, "ICs going on a 3-V diet," B. Prince and R. Salters, IEEE Spectrum, pp.22-25, May 1992.

[4] G.J. Hu et al., IEDM Tech. Dig., pp.710-713, Dec. 1982.

[5] L. Parrillo et al., IEDM Tech. Dig., pp.706-710, Dec. 1982.

[6] see chapter 3 of this thesis for example.

[7] see chapter 10 of "Operation and Modeling of The MOS Transistor," by Y.P. Tsividis, McGraw-Hill, 1987.

[8] J.Y.-C. Sun, C. Wong, Y. Taur and C.-H. Hsu, Digest of the Intl. Symposium on VLSI Technology, p.17, 1989.

[9] A.S. Grove, O. Leistiko and C.T. Sah, J. Appl. Phys., vol. 35, p.2695, 1964.

[10] F.K. Baker et al., IEDM Tech. Dig., pp.443-447, 1989.

[11] J.M. Sung et al., IEDM Tech. Dig., pp.447-451, 1989.

[12] A.J. Walker and P.H. Woerlee, pp.29-32, Proc. 19^{th} European Solid State Device Research Conference, Berlin, 1989.

[13] Y. Hiruta et al., IEDM Tech. Dig., pp.578-582, 1987.

[14] D.R. Wolters, Ph.D Thesis, University of Groningen, 1985.

[15] see chapter 2 of this thesis for example.

[16] I. Kato et al., IEDM Tech. Dig., pp.14-17, 1988.

[17] F. Matsuoka et al., IEDM Tech. Dig., pp.88-91, 1988.

[18] K. Kawabuchi et al., IEEE Trans. Elect. Dev., ED-32, p1685 (Sept 1985).

[19] F.K. Baker et al., IEDM Tech. Dig., pp.443-446, 1989.

[20] K. Oikawa et al., IEDM Tech. Dig., pp.79-82, 1991.

[21] H.G. Pomp et al., IEDM Tech. Dig., 1993.

Chapter 5

Silicon/Silicon Dioxide Barrier Height Adjustment

> *"Happily we all shoot at the moon with ineffectual arrows; our hopes are set on inaccessible El Dorado; we come to an end of nothing here below."*
>
> *R.L. Stevenson*

5.1 Introduction

In the constant pursuit of scaling devices to ever-decreasing dimensions and greater densities, another less obvious trend is taking place. This is the integration of other types of silicon devices onto the same slice of silicon as the normal N- and PMOS transistors. One important class of such devices is the nonvolatile memories. These are memory cells that can hold their memory state even after the supply voltage has been removed. Such devices are of great importance in a growing number of applications. For instance, in certain logic circuits known as microcontrollers, it is necessary for users to be able to program their own software into the chip for debugging purposes. The memory holding this program software has, up until now, been based on Electrically Programmable Read Only Memory (EPROM) which can be erased using ultra-violet light. A newer development is to use Flash Electrically Erasable and Programmable Read Only Memory (Flash EEPROM) [1] which adds flexibility over the EPROM technology in that the whole memory can quickly be erased with an electrical pulse. Other applications require that certain memory locations be constantly updated and nonvolatile. This requires the use of EEPROM which is a relatively large cell differing from the Flash version in that each byte of memory can be programmed and erased [2].

Operating EEPROM and Flash depend, in the majority of cases, on Fowler-Nordheim tunneling of electrons in thin oxides [3]. Due to the fact that this requires high voltages (greater than 13 V), integrated circuits containing scaled PMOS and NMOS transistors along with embedded EEPROM and Flash require special processes involving high voltage transistors with thicker gate dielectrics, deeper junctions and thicker isolation field oxides. Embedding even one bit of such nonvolatile memory would mean that the whole CMOS process must be changed to accommodate the required voltages for its programming and erasing. This is of course expensive so the thrust in embedding such devices into scaled integrated circuits is to lower these programming voltages. In the case of EPROMs, programming is achieved by hot electron injection at the drain side of the device [4]. A whole field of research has been opened to reduce the voltages necessary to program EPROMs with hot electrons. In [5], for example, the drain junction is very shallow and surrounded by an increased concentration of boron to increase the lateral fields in the device's channel region.

Figure 5.1 shows a schematic diagram of a Flash EPROM where programming is achieved by hot electron injection from the channel region and erasing by Fowler-Nordheim tunneling from the floating gate electrode [6].

In the case of EEPROMs and Flash where Fowler-Nordheim tunneling of electrons is concerned, reduction of program and erase voltages can be achieved by reducing the thickness of the tunnel dielectric. However, if this dielectric is too thin, the electrons on the floating gate may escape through direct tunneling to the substrate giving retention problems. In addition, the growth of very thin dielectrics uniformly and reproducibly is very difficult.

This chapter describes a new method to reduce voltages for Fowler-Nordheim electron tunneling that could be used to embed Flash and EEPROM devices into integrated circuits containing scaled NMOS and PMOS transistors [7]. It is based on the fact that ion implantation

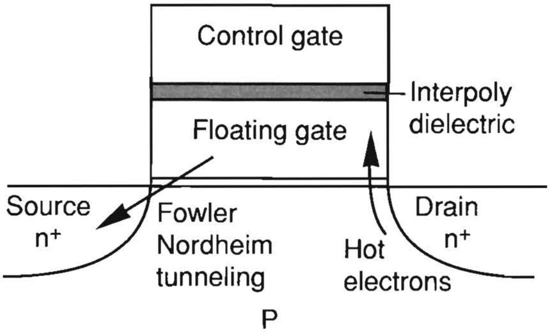


Figure 5.1. Schematic diagram of a Flash EPROM where programming is achieved by hot electron injection from the channel region and erasing by Fowler-Nordheim tunneling from the floating gate electrode.

into polycrystalline silicon covering an oxide layer increases the electron tunneling through the oxide layer. This can be seen in Fig. 5.2 which shows the tunnel current in MOS capacitors with 15 nm of gate oxide dielectric. The control capacitor is the normal MOS capacitor with polycrystalline silicon gate electrode. The other two capacitors have undergone either arsenic or germanium implantation into the gate polycrystalline silicon gate. The fluence and energy of these implants were the same in both cases. An interesting point to note is that the shifts in tunneling to smaller voltages are almost identical for the arsenic and germanium cases.

Since both ions have very similar atomic masses, this suggests that the effect is due to the physical damage of the implantation itself rather than any chemical effect of either the arsenic or germanium. The implantation of silicon into gate oxides has been used elsewhere [8] but is based on a different idea, namely that the implanted silicon forms traps in the gate oxide which can be used to store electrons and thus act as a nonvolatile memory element.

The chapter begins with a summary of Fowler-Nordheim tunneling along with a literature survey of previous attempts to decrease such voltages. Next, the new experiments, carried out on simple capacitors, micron-size transistors and Flash EEPROMs, are described followed by results and discussion. After that, a summary is given including the importance of such effects in the drive to "systems on silicon".

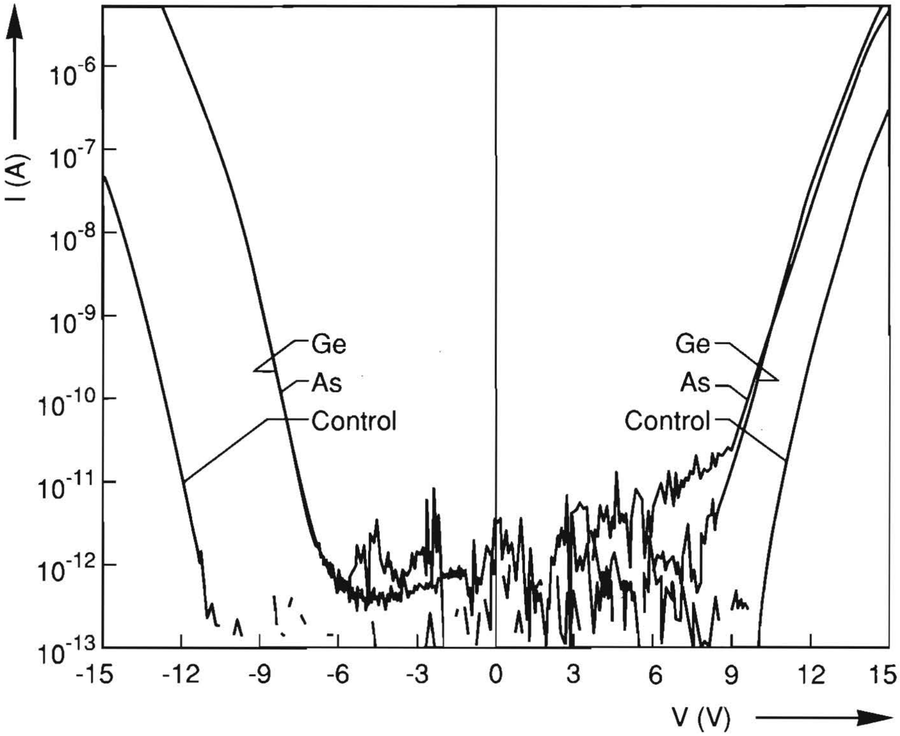


Figure 5.2. Enhanced tunneling in MOS capacitors due to ion implantation in the polycrystalline silicon gate electrode covering 15 nm of thermal oxide. The fluence and energy were 10^{15} $/cm^2$ and 80 keV respectively for both the arsenic and germanium. The implanted polycrystalline silicon was 100 nm thick.

5.2 Fowler-Nordheim Tunneling

Figure 5.3 gives an energy-band diagram of the tunneling process in which an electron tunnels from the conduction band of the silicon to the conduction band of the oxide and subsequently flows towards the metal anode.

The equation describing this process can be expressed as

$$J = AE_{ox}^2 exp(-B/E_{ox}) \tag{5.1}$$

where J is the electron tunnel current density, E_{ox} is the field in the oxide, and A and B are constants. In particular, A = $3 \times 10^6/\Phi$ (A/cm^2) and B = $48.3\Phi^{3/2}$ (MV/cm) where Φ is the barrier energy in

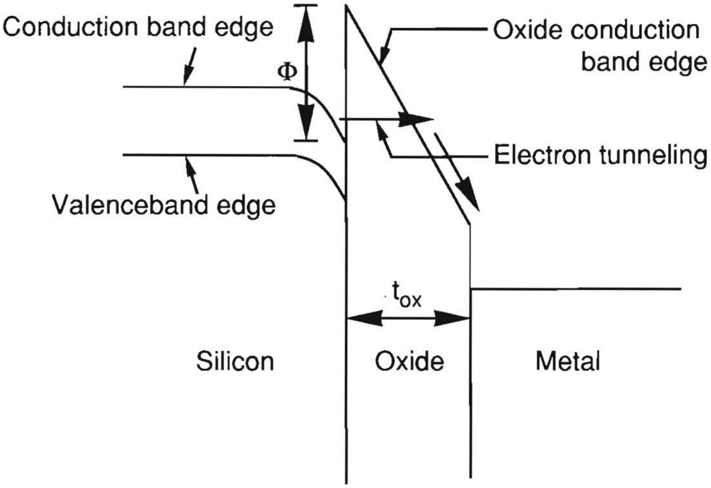


Figure 5.3. Energy-band diagram of the tunneling process in which an electron tunnels from the conduction band of the silicon to the conduction band of the oxide and subsequently flows towards the metal anode.

eV. If the oxide is too thin, the electrons can tunnel directly from the cathode to the anode without appearing in the oxide conduction band. In the cases studied here, the oxide is always thick enough to present a triangular barrier to the electron without direct tunneling (see Fig. 5.3).

The ability to enhance electron tunneling through silicon dioxide has been reported in various studies. In [9], tunneling through chemically vapour-deposited oxides was increased by making the oxide's interfaces silicon-rich. By modifying these interfaces, the desired current-voltage characteristics could be obtained.

In [10], the effect of arsenic implantation through gate oxide and the growth of gate oxide on implanted substrates were reported. Above a certain arsenic implantation dose, the Fowler-Nordheim constants, A and B, reduced. The effect of an arsenic implantation into a patterned polycrystalline silicon gate electrode with gate edges on thin oxide was shown to result in a kinked Fowler-Nordheim plot. These effects were thought to occur due to the presence of trapping centres in the oxide due to the implantation or the oxidation of implanted silicon.

The fact that oxides grown on heavily doped silicon substrates have smaller tunnel barrier heights than those grown on lightly doped substrates was used in the fabrication of EEPROMs [11] with low voltage

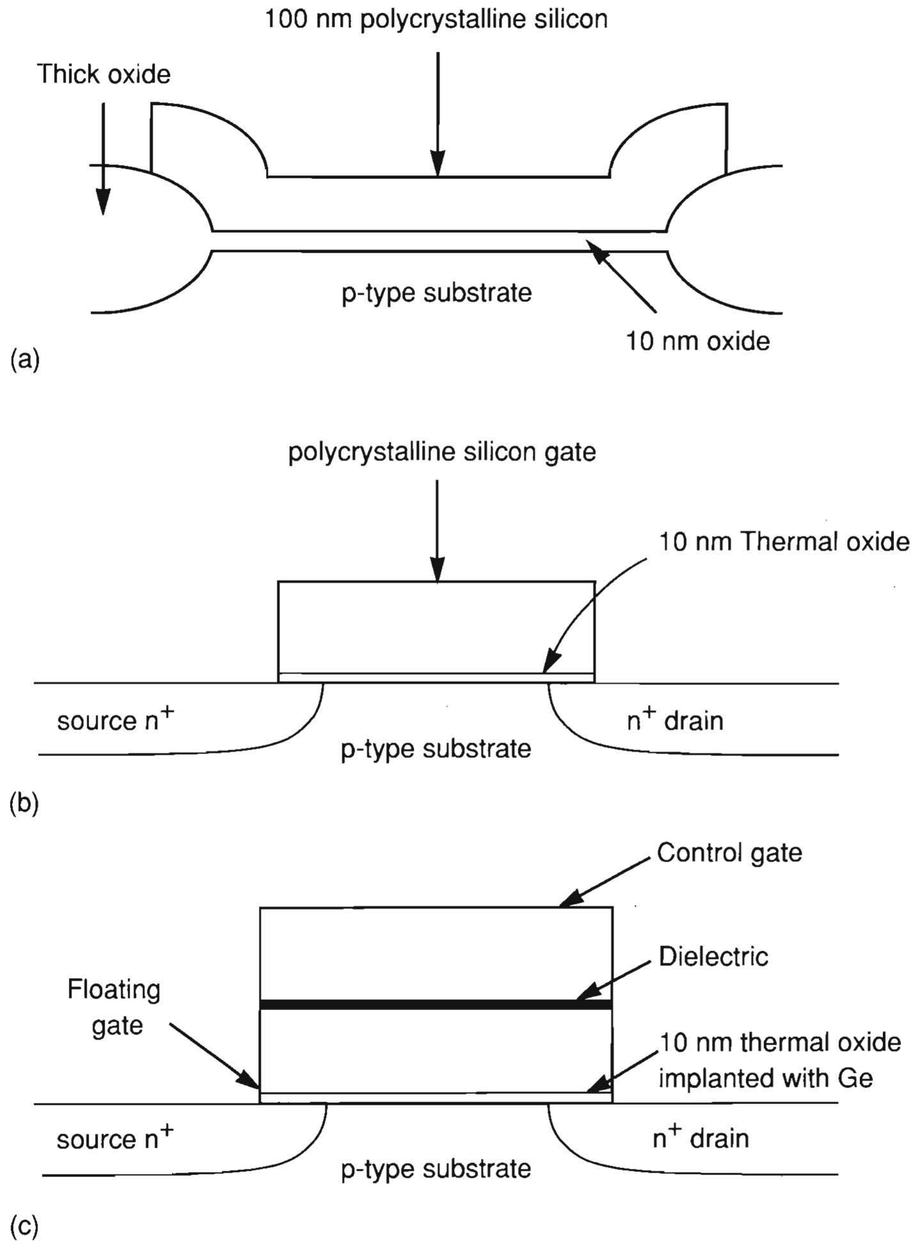


Figure 5.4. Schematic diagrams of devices used to study enhanced tunneling. (a) Capacitor with area of $6 \times 10^{-3} cm^2$ and oxide thickness of 10 nm. (b) MOS transistor with 10 nm gate oxide and gate length and width of 1μm and 1.5μm respectively. (c) Flash EEPROM with gate oxide of 10 nm, gate length and width of 1μm and 1.5μm respectively, and interpolycrystalline dielectric of 50 nm silicon oxynitride.

program and erase.

In all above cases, no commercial devices have ever reached the marketplace possibly because of manufacturing difficulties. This chapter describes the use of an effect in which ion implantation into polycrystalline silicon covering an oxide layer increases the electron tunneling through the oxide layer [12]. The implantation leads to atomic mixing in the implanted target due to interactions with both the primary ions and the secondary recoils [13]. As a result, the oxide interfaces are no longer sharp. The ion beam mixing also leads to changes in the bulk of the oxide which can have important consequences for charge transport.

This chapter presents a study of enhanced tunneling using ion beam mixing carried out using germanium (Ge^+) implantation in metal-oxide-silicon (MOS) structures. Ge^+ is useful because it is a relatively heavy ion allowing implant damage to be caused at low implant doses. In addition, relatively steep implant and well defined damage profiles can be obtained compared to lighter ions such as silicon. Furthermore, it does not dope the underlying silicon substrate, unlike arsenic, and thus does not affect MOS electrical parameters such as threshold voltage.

Experimental results, to be presented here, suggest that the implantation not only leads to oxide interface barrier reduction but also bulk oxide damage which eases electron transport. This was done by fabricating a Flash EEPROM which is ideal to study low current effects in thin oxides. Despite this electron transport, it will be shown that Flash EEPROMs, programmable and erasable at low voltages due to this technique, have sufficient charge retention for possible use as non-volatile memory. This implies that, besides thinner tunnel dielectrics, this provides an alternative path to make low voltage Flash EEPROMs. Such devices, with their thicker tunnel oxides, experience less electric field stresses and have less parasitic floating gate-to-substrate capacitances. The latter advantage means that capacitive coupling between the control gate and floating gate is enhanced with the result that a larger fraction of the program voltage is available for writing and erasing the device.

5.3 Experiment

Silicon substrates used in this study were $< 100 >$ p-type wafers with an epitaxial top layer of 5 μm thickness and 10 Ωcm resistivity.

After 10 nm thermal oxide growth for the tunnel dielectric of the capacitors, transistors and Flash EEPROM, 100 nm of polycrystalline silicon was deposited and doped with arsenic or phosphorus implantation. The ion beam mixing implantation was carried out using germanium (Ge^+) at a dose of $10^{15} cm^{-2}$ and energy of 80 keV which ensures that the gate oxide is situated in the tail of the implantation. This was implanted over the whole wafer. The polycrystalline silicon was then patterned using photolithography and plasma etching. Standard processing techniques were followed to complete the fabrication. Included in this are standard annealing steps at about 900^{o}C. The effect of different anneal steps was not studied. Also, the sensitivity of tunnel enhancement to implant dose and energy and to variations in polysilicon thickness was not studied but would certainly be important to study manufacturability of this process.

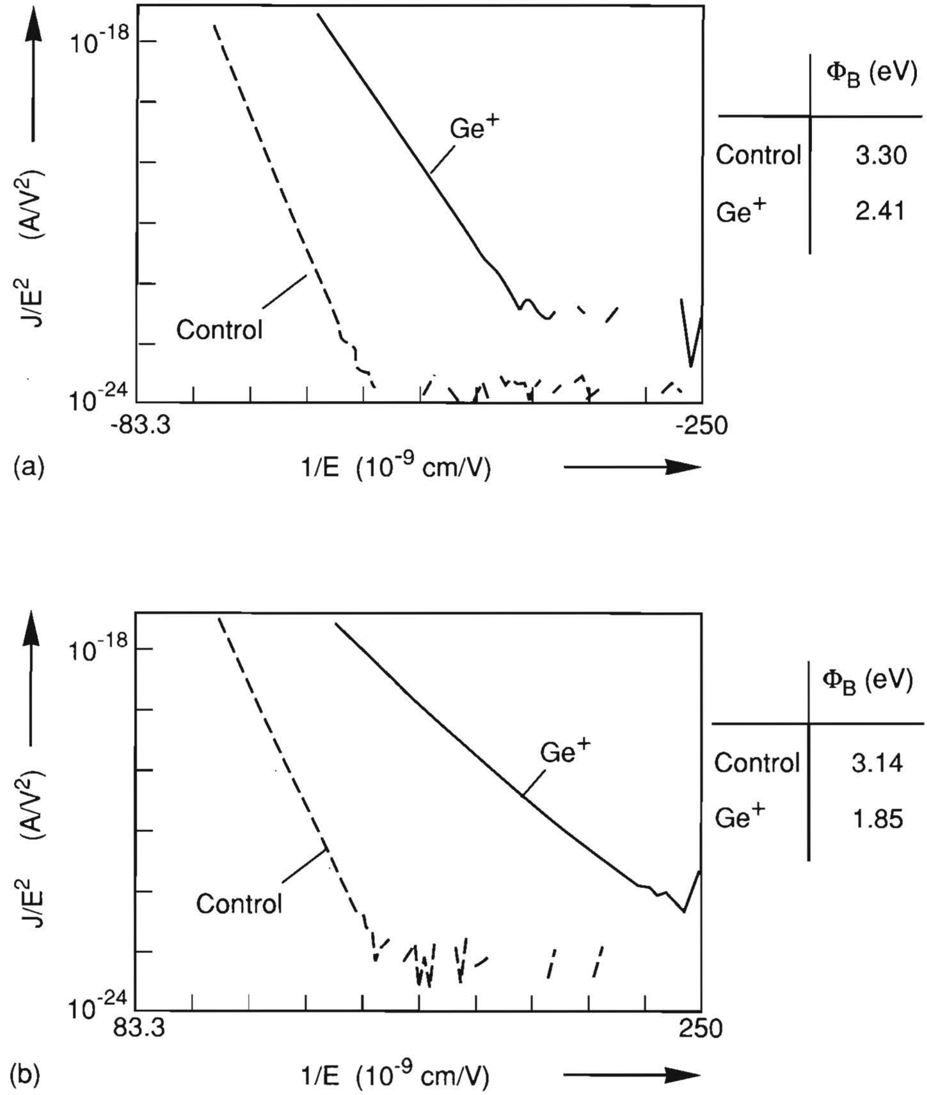
Figure 5.4 shows schematic diagrams of the three types of fabricated devices.

5.4 Results and Discussion

Figure 5.5 shows Fowler-Nordheim plots for electron tunneling in the capacitors indicating the barrier reduction caused by the Ge^+ implantation. The barrier height to electron tunneling is measured from the slope of the line in the Fowler-Nordheim plot (see for instance [10]).

From a large population (> 100) of such capacitors, the charge-to-breakdown, Q_{bd}, was measured by forcing a constant current of density 10^{-2} A/cm^2 through the gate oxide. It was seen that the implanted devices had a Q_{bd} between 1 and 10 C/cm^2 which compares favorably with the unimplanted devices. This is an important parameter since electrons traverse the tunnel oxide during programming in EEPROM and Flash devices. A simple calculation can be made to explain this point. Consider an EEPROM where the tunnel oxide region has an area of 1 μm^2. The floating gate capacitance is in the order of 1 fF. Therefore, for a threshold voltage swing of 5 volts, the amount of charge traversing the tunnel oxide is 5 fC. For a device that has to switch 10^5 times, the total amount of charge per unit area crossing the tunnel oxide is then 0.1 C/cm^2.

Figures 5.6 and 5.7 show that an increased field is necessary to force a constant current through the implanted capacitors (in this case



	Φ_B (eV)
Control	3.30
Ge$^+$	2.41

	Φ_B (eV)
Control	3.14
Ge$^+$	1.85

Figure 5.5. Fowler-Nordheim plots of electron tunneling in implanted and unimplanted capacitors. (a) Tunneling from gate to substrate. (b) Tunneling from substrate to gate. Gate area is $6 \times 10^{-3} cm^2$ and Ge^+ implantation is $10^{15} cm^{-2}$ at 80 keV. Also shown are tunneling barrier heights, Φ_B. The control device has no Ge^+ implantation.

with 17.5 nm gate oxide thickness). The increased electron trapping, causing this effect, may be due to the oxide being silicon-rich [16]. The significance of this is that trapping must not be allowed to become too large. Otherwise, it becomes increasingly more difficult to tunnel electrons through the thin oxide to and from the floating gate of the EEPROM or Flash. As will be seen later, this does not seem to pose any problem.

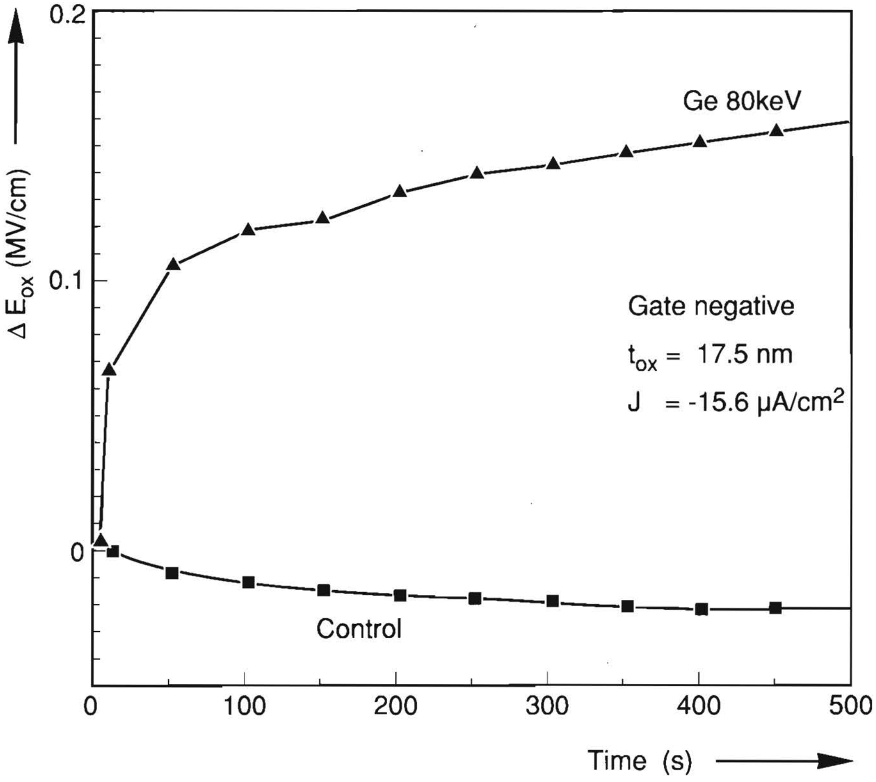


Figure 5.6. Change in oxide field for a constant current stress with negative capacitor gate.

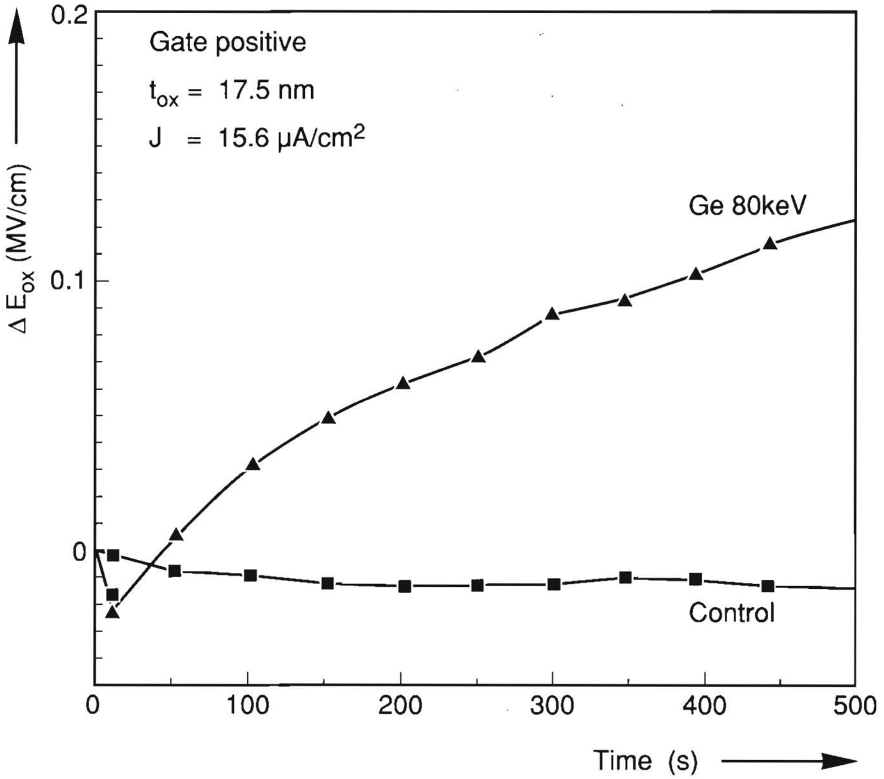


Figure 5.7. Change in oxide field for a constant current stress with positive capacitor gate.

Figure 5.8 shows source-drain and gate currents in the MOS transistors. The increase in gate current from substrate to gate and from gate to substrate due to the Ge^+ implantation is clearly seen. This is entirely due to Fowler-Nordheim tunneling of electrons and not to hot electron injection since the drain voltage is too low for such a process to take place. The slight increase in threshold voltage of the germanium implanted device is well within the variations in the control threshold voltage measured in a large population of devices.

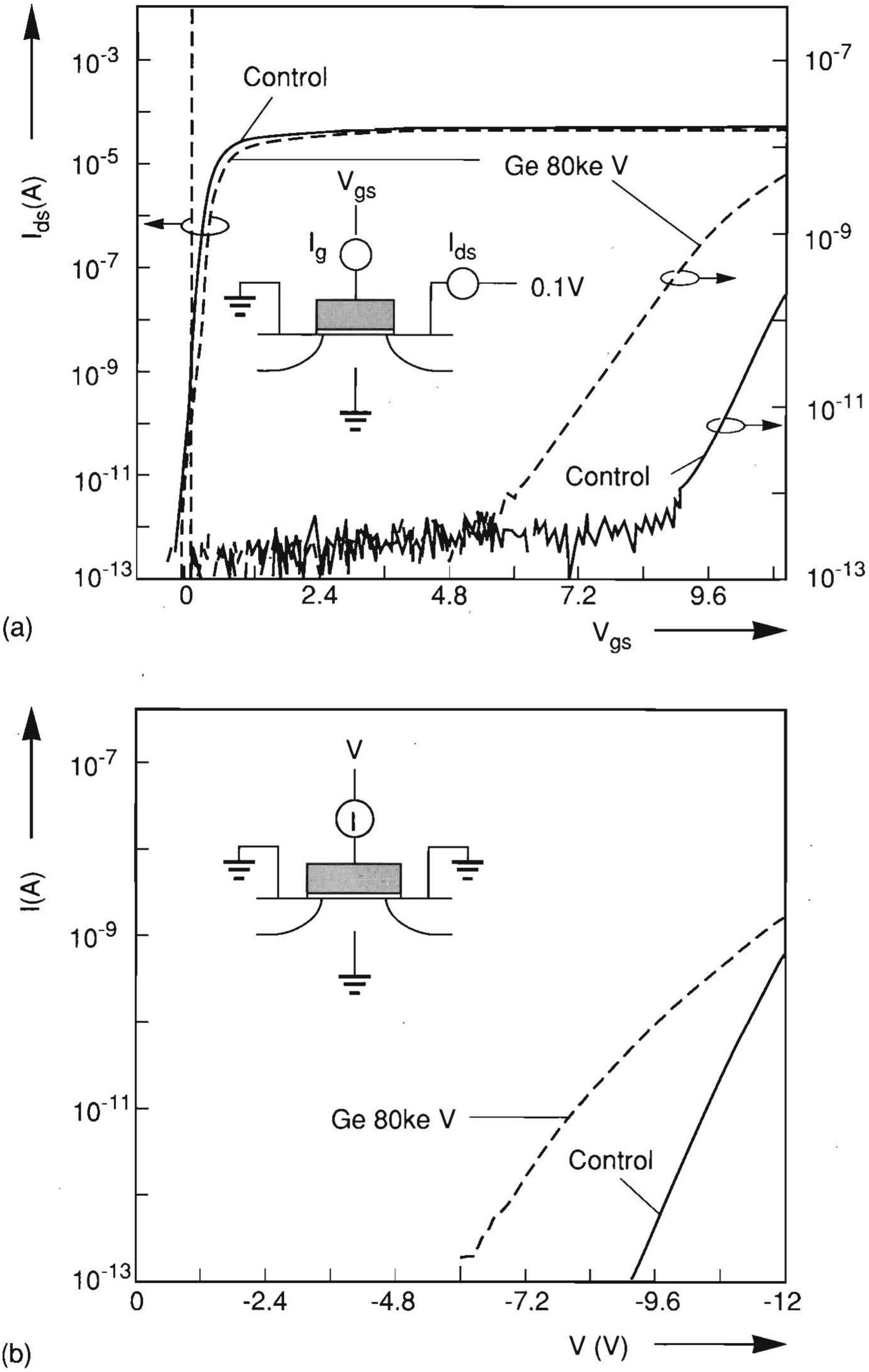


Figure 5.8. MOS transistor characteristics for a device with gate length and width of 1μm and 1.5μm respectively. (a) Source-drain and gate currents plotted as a function of gate voltage. (b) Gate current due to electron tunneling from the gate electrode.

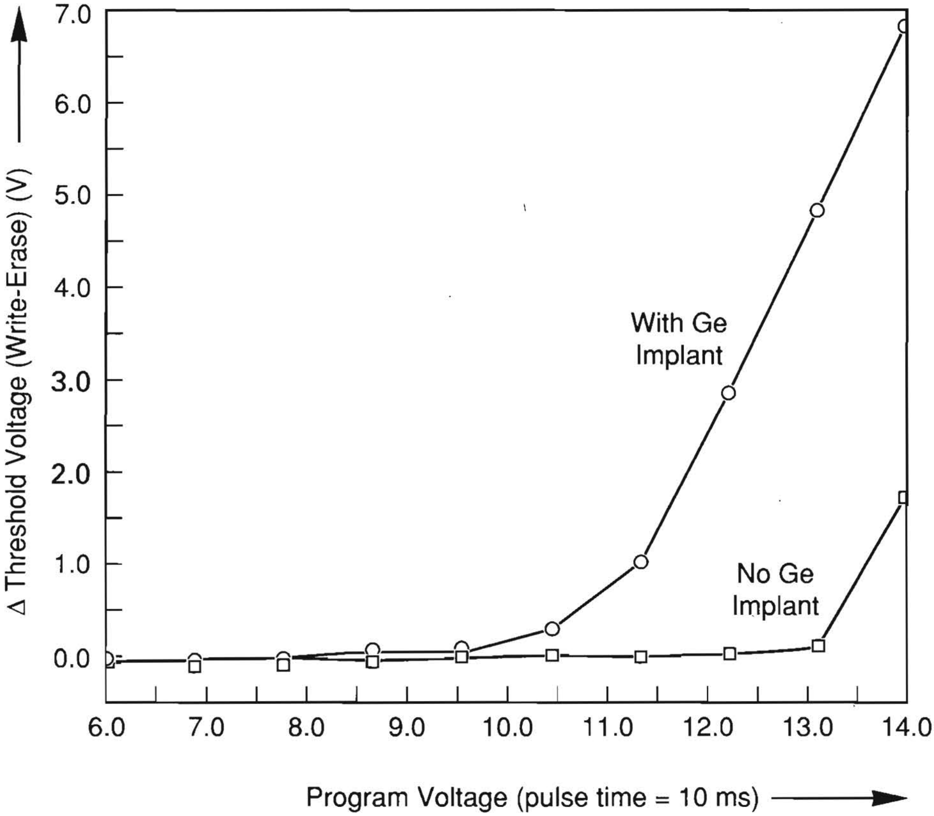


Figure 5.9. Flash EEPROM program characteristics. ΔV_T is the difference in threshold voltage between written and erased states. Ge^+ implantation dose is $10^{15} cm^{-2}$.

Figure 5.9 shows programming characteristics of the Flash EEPROM. To tunnel electrons to the floating gate, the source, drain and substrate are held at ground while the control gate of the device is pulsed to the program voltage for 10 ms. Through capacitive coupling about 60% of the control gate voltage is dropped across the 10 nm injector oxide leading to tunneling. This can be appreciated if the control gate - floating gate - drain is considered as a series connection of two capacitors with the floating gate as common electrode. The system operates simply as a voltage divider with 60% of the control gate voltage dropped across the tunnel oxide.

Similarly, to tunnel electrons from the floating gate to the drain, the control gate is grounded while the drain is pulsed with the program voltage. The source and substrate are left floating. The amount of charge on the floating gate affects the voltage required on the control gate to form an inversion layer in the silicon substrate. This control gate voltage is known as the threshold voltage. The more electrons there are on the floating gate, the larger the control gate voltage needs to be to invert the silicon surface. After forcing electrons from the floating gate, the floating gate can be left with a net positive charge resulting in a lower control gate voltage for silicon inversion. The difference in threshold voltage between these two states for the Flash EEPROM cell is plotted as a function of the program voltage in Fig. 5.9. The tunnel current enhancement due to ion beam mixing leads to the much larger threshold voltage differences shown.

Figure 5.10 shows retention characteristics at 85°C for the Flash EEPROM cells with Ge^+ implantation. Besides being a common chip environment temperature, 85°C is the normal temperature at which retention is tested in metal-nitride-oxide-silicon nonvolatile memories which show similar characteristics [14]. The device was pulsed either to a positive or negative threshold voltage with the subsequent threshold voltage being monitored with time. The control device without Ge^+ implantation is not shown since no change was seen in the threshold voltages within the measured time shown. The logarithmic discharge process is similar to that measured in metal-nitride-oxide-silicon nonvolatile memories [15]. In the latter, the process is based on direct tunneling of electrons from traps in the nitride layer to the silicon. In the case of the Ge^+ implanted Flash EEPROMs, primary ion damage and knock-on recoil implantation of silicon give rise to the presence of implant-induced traps [16]. It is envisaged that electrons tunnel to and from the floating gate via these traps. The logarithmic nature of the discharge process suggests that such devices could be used as nonvolatile memories which are programmable and erasable at low voltages and have sufficient charge retention for several years operation. In most applications involving nonvolatile memories, a guarantee of 10 years retention is usually given. In many applications, this is many times what is actually required. A retention of a couple of years is sufficient in many cases.

Figure 5.11 shows the endurance of an implanted Flash device. Up

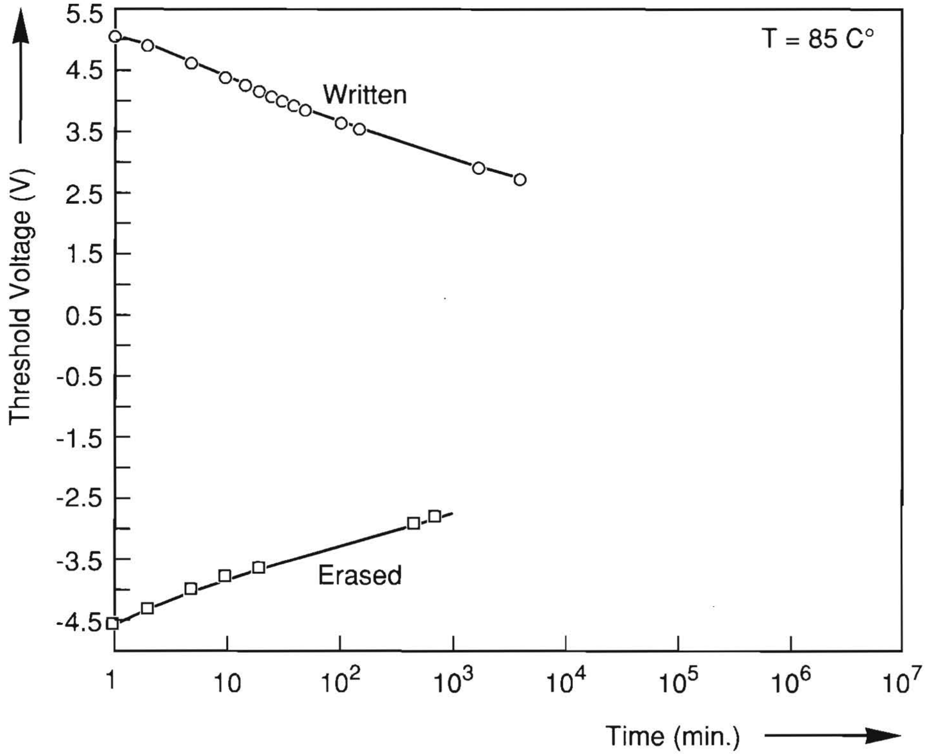


Figure 5.10. Flash EEPROM charge retention characteristics at 85°C for the implanted device showing threshold voltage versus time after programming or erasing. Implantation dose is $10^{15} cm^{-2}$. The threshold voltage of the unimplanted control device did not change in the measured time shown.

to 5×10^5 pulses of 15 volts and 2 ms were applied with the threshold voltage being measured after certain numbers of cycles. Clearly, the implanted flash device can be cycled many times and still has a large enough threshold voltage window.

5.5 Summary

MOS structures were fabricated to investigate enhanced Fowler-Nordheim tunneling in thin oxides due to ion beam mixing. Ions of Germanium were implanted into a 100 nm polycrystalline silicon layer deposited on 10 nm thermal silicon dioxide such that the tail of the implant profile contains the thin oxide. Besides simple MOS capaci-

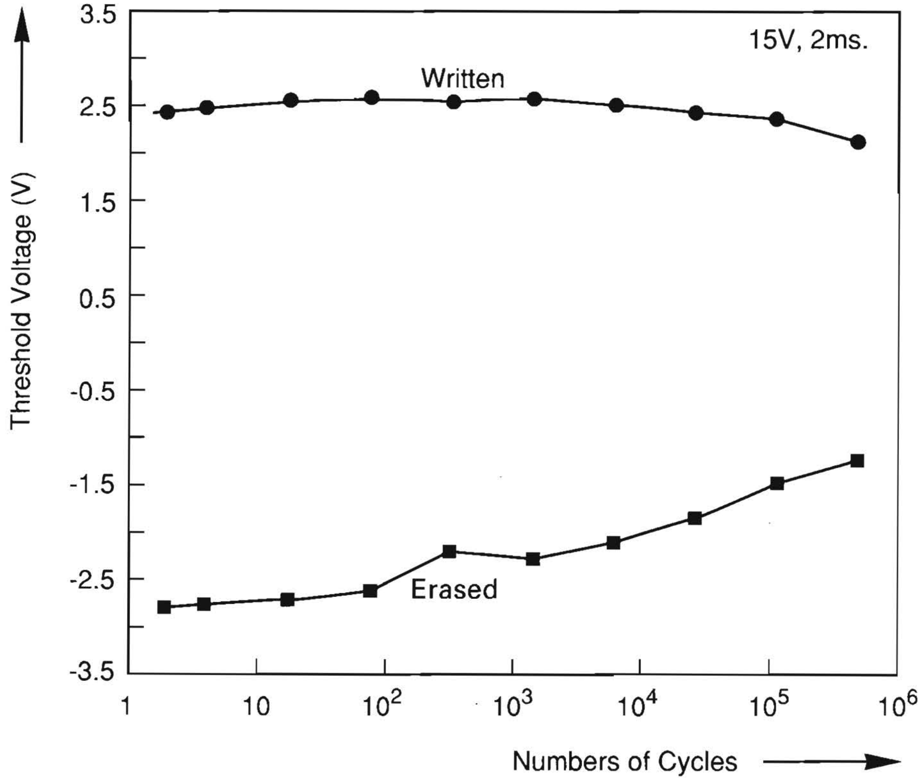


Figure 5.11. Flash EEPROM cycle endurance characteristics with pulses of 15 volts and 2 ms.

tors and transistors, Flash EEPROM cells were fabricated for the first time using this technique. Using $1 \times 10^{15} cm^{-2}$ at 80 keV the Fowler-Nordheim tunneling barrier reduced by about 0.9 eV at the polycrystalline gate/oxide interface and by 1.3 eV at the oxide/substrate interface. The consequently lower program/erase voltages in the Flash EEPROM were measured. Flash EEPROM charge retention measurements show that the discharge process is logarithmic in time. This leads to the possibility of Flash EEPROMs being programmed and erased at low voltages and having sufficient charge retention for several years operation.

As has been stated above, the effect of variations in germanium energy and dose was not studied but is certainly important for any future study into the manufacturability of this process.

The drive to smaller dimensions in integrated circuits has been fueled by the requirement of increased functionality per unit cost. Besides by cramming more transistors onto a given silicon area, increased functionality can also be achieved by integrating different types of devices onto the same slice of silicon. In the case of nonvolatile memories such as EPROMs, EEPROMs and Flash, relatively high voltages are required for their operation. This clashes with the need for low voltages for the scaled NMOS and PMOS transistors occupying the rest of the silicon. This chapter has shown a possible path to lower voltage operation of those nonvolatile memories that rely on Fowler-Nordheim electron tunneling for their operation, such as EEPROMs and Flash devices making it possible to combine such devices with the scaled transistors and thus to increase the functionality of an integrated circuit.

5.6 References

[1] C. Kuo, M. Weidner, T. Toms, H. Choe, K.M. Chang, A. Harwood, J. Jelemensky and P. Smith, IEEE J. of Solid State Circuits, vol.27, pp.574-582, April 1992.

[2] For a good summary of the different types of nonvolatile memory, see "Nonvolatile Semiconductor Memories" ed. C. Hu, pp.5-9, article "A Memory to Remember", by M. Bloom, IEEE Press, New York, 1991.

[3] M. Lenzlinger and E.H. Snow, J. Appl. Physics, vol.40, pp.278-283, 1969.

[4] For an overview of the EPROM, see "Nonvolatile Semiconductor Memories" ed. C. Hu, pp.59-62, article "An EPROM's integrity starts with its cell structure", by M. Woods, IEEE Press, New York, 1991.

[5] Y. Ohshima et al., IEDM Tech. Dig., pp.95-98, 1990.

[6] "Endurance Brightens the Future of Flash", by K. Robinson, in "Nonvolatile Semiconductor Memories" ed. C. Hu, pp.167-169, IEEE Press, New York, 1991.

[7] A.J. Walker and J. Politiek, Appl. Phys. Lett., vol.63, pp.758-760, 1993.

[8] T. Hori, T. Ohzone, Y. Odake and J. Hirase, IEDM Tech. Dig., pp.469-472, 1992.

[9] D.J. DiMaria and D.W. Dong, Appl. Phys. Lett, vol.37, pp.61-64, 1980.

[10] H.L. Peek and D.R. Wolters, Proc. SSDM, pp.487-490, 1986.

[11] H. Nozawa, N. Matsukawa and S. Morita, IEEE Trans. Elec. Dev., ED-33, pp.275-281, 1986.

[12] M.G. Stinson and C.M. Osburn, J. Appl. Phys., 67, p4190, 1990.

[13] see for instance chapter 7 in "Ion Implantation and Beam Processing," edited by J.S. Williams and J.M. Poate, Academic Press, Sydney 1984.

[14] S. Minami and Y. Kamigaki, IEEE Trans. Elec Dev., ED-38, pp. 2519-2526, Nov. 1991.

[15] L. Lundkvist, I. Lundstrom and C. Svensson, Solid-State Elec., vol.16, pp.811-823, 1973.

[16] A. Kalnitsky, A.R. Boothroyd and J.P. Ellul, IEDM Tech. Dig. 1988, pp. 516-519.

Summary

"Little do ye know your own blessedness; for to travel hopefully is a better thing than to arrive, and the true success is to labour."
R.L. Stevenson

In the last thirty years or so, many engineers and scientists have travelled hopefully towards reliable integrated circuits containing scaled NMOS and PMOS transistors. This thesis has shown one such path.

The journey started with the modeling of the electron and hole mobility in the inversion layer of MOS devices. A semi-empirical model was used to describe the mobility as a function of an effective normal field which is one of the most important parameters in such a formulation. This is important to be able to model scaled devices in circuit simulators and so to predict behaviour in increasingly complex systems.

The next stop on the way was the important matter of making shallow boron junctions for scaled PMOS devices. This involved the preamorphization of the source and drain regions before boron implantation. It was concluded that, despite shallow junctions having been made, such a technique would be very difficult to control in a manufacturing environment due to the contraints imposed by the temperature budget and resultant defects. It was seen that a simple low energy boron difluoride implantation followed by a low temperature anneal resulted in ideal shallow junctions. This topic remains extremely important with research into areas of solid source diffusion, gas phase doping and low temperature annealing.

The difficulty in making scaled PMOS devices was emphasized in the next stop on the journey when boron-doped polysilicon gates were studied. It was concluded that such a process leads to stable devices provided the temperature budget is reduced to prevent boron diffusion to the channel region. Furthermore, it was shown that the p-type gate instability is independent of boron being present in the gate and gate dielectric but depends only on the work function of the gate electrode. Hot-carrier resistant devices were made using the process described.

With thinner gate oxides and shallower junctions, integrated circuits containing scaled devices cannot accommodate large voltages. This becomes a problem when other types of devices are integrated along with the scaled NMOS and PMOS transistors. One important class of devices is the nonvolatile memories which usually require voltages around 13 volts for programming and erasing. The final stop on this journey dealt with a possible solution to lowering such voltages in nonvolatile memories which depend on Fowler-Nordheim tunneling for programming or erasing. It was seen that such voltages could be reduced by implanting a heavy ion into a thin polysilicon layer on top of the tunnel dielectric. This resulted in enhanced tunneling in Flash devices which can retain their charge for a several years.

After surveying these stops along the journey, there is finally enough time to look to the future. As always, the shape of the future can be discerned from studying the present and past.

The past few years have witnessed the increasing importance of BiCMOS which combines the advantages of MOS and bipolar transistors on the same slice of silicon. The synergy afforded by their combination makes it worthwhile to contend with the resultant process complexity.

Another important development is the growing demand for memories embedded along with logic circuits. In particular, the availability of nonvolatile memories "on-board" logic circuits is becoming of vital importance. Again, the complex processes needed to integrate such memories are becoming cost-effective to run as demand increases for the resultant chip's increased functionality.

The combination of digital and analogue elements on the same chip can also be discerned from the need to combine switched capacitors with normal scaled MOS transistors.

From these developments we can see the beginnings of an important manoeuvre to increase functionality. The classic path has always been to shrink dimensions of transistors to squeeze more logic gates onto every square millimetre of silicon. The required developments in process, design and equipment technology have all been driven by users' demands for the increased functionality. This path will remain. The important new front to increased functionality consists of integrating different types of devices onto the same slice of silicon. Again, the necessary developments in process, design and equipment technology will be pushed by demands for the increased functionality. At the moment, all such devices can be made using existing materials in normal integrated circuit processes but the time will come when new materials will offer such great advantages that they too will be integrated along with the normal silicon devices. This will then offer possibilities not only for improved performance of existing devices but also for totally new types of devices which will, in turn, be integrated "on-board". The spoils will go to those companies that are able to master the complexities of these veritable "systems-on-silicon".

Samenvatting

Gedurende de afgelopen dertig jaar zijn vele technologen en wetenschappers op zoek geweest naar wegen, waarlangs betrouwbare geintegreerde circuits met geschaalde NMOS en PMOS transistoren gemaakt kunnen worden. Dit proefschrift beschrijft zo'n weg.

De eerste halte op deze zoektocht was het modelleren van de mobiliteit van electronen en gaten in de inversielaag van MOS devices. Door middel van een semi-empirisch model is deze mobiliteit beschreven als functie van een effectief normaalveld, een van de belangrijkste parameters in zo'n formulering. Het belang hiervan ligt in de mogelijkheid hiermee geschaalde devices in circuitsimulatoren te modelleren en aldus een voorspelling te doen over het gedrag in toenemend complexe systemen.

Als volgende halte werd het maken van ondiepe boorjuncties voor geschaalde PMOS devices behandeld, een onderwerp van groot belang. Dit hield een pre-amorfisatie in van de source- en draingebieden, voorafgaande aan de boorimplantatie. Geconcludeerd werd dat, ondanks dat op deze manier inderdaad ondiepe juncties zijn te maken, in een produktie omgeving een dergelijke techniek zeer moeilijk te controleren zal zijn vanwege de beperkingen, die het temperatuurbudget en de resulterende defecten opleggen. Het bleek dat een simpele lage energie boordifluoride implantatie, gevolgd door een lage temperatuur anneal, in ideale ondiepe juncties resulteerde. Dit onderwerp blijft zeer relevant voor onderzoek op het gebied van solid source diffusie, gasfase dotering en lage temperatuur annealen.

Vervolgens werd stilgestaan bij boorgedoteerde polysilicium gates, en werden de problemen bij het maken van geschaalde PMOS devices benadrukt. De conclusie werd getrokken dat een dergelijk proces tot stabiele devices leidt, onder voorwaarde dat het temperatuurbudget gereduceerd wordt, om boordiffusie naar het kanaal te vermijden. Bovendien werd aangetoond dat de p-type gate instabiliteit niet beinvloedt wordt door de aanwezigheid van boor in de gate en in het gate dielectricum, maar alleen afhangt van de werkfunctie van de gate electrode. Via het beschreven proces zijn devices gemaakt, die bestand zijn tegen hete ladingsdragers.

Naarmate gate oxides dunner en juncties ondieper worden, kunnen geintegreerde circuits met geschaalde devices steeds minder hoge spanningen verdragen. Dit is een probleem zodra niet alleen geschaalde NMOS en PMOS transistoren geintegreerd worden, maar ook andere types devices. Een belangrijke klasse devices vormen de niet-vluchtige geheugens, die in het algemeen een spanning van ongeveer 13 volt benodigen voor programmeren en wissen. De laatste halte op onze zoektocht behandelt een mogelijke oplossing om deze spanning te verlagen. Deze oplossing is gebaseerd op de toepassing van Fowler-Nordheim tunneling voor programmeren en wissen. Programmeer- en wisspanningen bleken gereduceerd te kunnen worden door zware ionen te implanteren in een dunne polysiliciumlaag op het tunneldielectricum. Dit resulteert in versterkt tunnelen in Flash devices. Deze devices kunnen hun lading enkele jaren behouden.

Na het beschouwen van de verschillende haltes op onze zoektocht, kunnen we nu de tijd nemen om naar de toekomst te kijken. Zoals altijd kan de toekomst voorspeld worden door het heden en verleden te bestuderen en te interpreteren.

De laatste jaren hebben een toenemende interesse laten zien in BICMOS, een techniek die de voordelen van MOS en bipolaire transistoren op eenzelfde siliciumplak combineert. Door de synergie van deze combinatie is het lonend het produktieproces, met de daarbij behorende toegenomen procescomplexiteit, onder de knie te krijgen.

Een verdere belangrijke ontwikkeling is de toename in het belang van geheugens ingebed in logische circuits. In het bijzonder de mogelijkheid om niet-vluchtige geheugens op deze manier te combineren met logische circuits wordt van essentieel belang. Ook hier wordt het lonend de complexe processen, nodig om zulke geheugens in te passen, voor lief te nemen naarmate de vraag naar chips met een resulterende grotere functionaliteit toeneemt.

Ook de trend om digitale en analoge elementen op dezelfde chip te combineren kan afgeleid worden uit de behoefte geschakelde capaciteiten met normale, geschaalde MOS transistoren te combineren.

Deze ontwikkelingen wijzen erop, dat een toename van de functionaliteit steeds belangrijker wordt. De klassieke weg is altijd geweest om de dimensies van transistoren te verkleinen, om aldus meer logische poorten op elke vierkante millimeter silicium te "persen". De hiervoor benodigde ontwikkelingen op het gebied van procestechnologie, design en apparatuur zijn steeds in gang gezet, doordat de gebruikers van IC's behoefte hadden aan een grotere functionaliteit. Deze behoefte zal ook in de toekomst een drijvende kracht zijn achter nieuwe ontwikkelingen. Een nieuwe, belangrijke manier om een toename in functionaliteit te realiseren is het integreren van verschillende types devices op eenzelfde siliciumplak. Ook hier zal de behoefte aan een toename in functionaliteit een belangrijke stimulans zijn voor de benodigde ontwikkelingen op het gebied van procestechnologie, design en apparatuur. Op dit moment kunnen alle devices gemaakt worden met bestaande materialen in normale IC processen, maar in de toekomst zullen nieuwe materialen dusdanig grote voordelen bieden, dat ook deze met de normale siliciumdevices geintegreerd zullen worden. Dit opent de deur zowel naar betere prestaties van de bestaande devices, als naar volledig nieuwe types van devices, die op hun beurt geintegreerd zullen worden. De grootste winst hiermee zal gemaakt worden door die firma's die de ingewikkelde processen beheersen, nodig om deze "systems-on-silicon" te realiseren.

Nawoord

The research described in this thesis was carried out at the Philips Research Laboratories, Eindhoven as part of the research programme. I would like to express my thanks to the management, especially Max Collet and Fred van Ommen, for allowing me to pursue this research and the resulting degree. My gratitude goes to Frans Klaassen for his guidance throughout, Pierre Woerlee and Herbert Lifka for their support from my first day in the Netherlands, Stefan Louwers for assistance with the Dutch language, Peer Zalm for his many SIMS analyses, Ann de Veirman for expert TEM cross sections, Joep Baggerman for his expertise with Latex, Dim Wolters and Adrie Zegers-van Duijnhoven for measurements and discussions, members of the MOS technology research group for support, and of course all members of FABWAG who actually made the devices. My thanks goes to the Audio Visuele Dienst in the Research Laboratory, in particular to Mieke Spaan and Hennie Herps for their expert assistance. Finally, I would like to thank Luciane for all her support throughout.

Andrew J. Walker

Eindhoven, 7 december 1993.

Curriculum Vitae

Andrew Walker was born in Sialkot, Pakistan on the 11^{th} of December, 1962. After secondary education at Blairgowrie High School, Perthshire, Scotland, he went on to study physics at The University of Dundee, Scotland graduating with First Class Honours in 1985. In October 1985, he joined the Philips Research Laboratories in Eindhoven. His initial research involved the technology of deep submicron CMOS devices which resulted in several publications and conference presentations. From 1990, his research has involved the technology of nonvolatile memories. Besides several publications, he also has several patent applications.

Stellingen behorende bij het proefschrift:

THE PHYSICS AND TECHNOLOGY OF SUBMICRON MOS DEVICES

A.J. Walker
dinsdag 8 februari 1994

TU Eindhoven

1. The introduction of fluorine by implantation of boron difluoride in preamorphized silicon layers leads to the almost complete absence of boron diffusion during rapid thermal anneals at temperatures at and above 1000°C in nitrogen.

   This thesis, chapter 3.

2. Ion beam mixing of the silicon/silicon dioxide interface leads to enhanced Fowler-Nordheim tunneling of electrons through the oxide.

   This thesis, chapter 5.

3. From many years of experience in the two countries, it can be said that the preoccupation with money is greater in the Netherlands than in Scotland. The myth of the frugal Scotsman however will no doubt continue.

4. The economic development of the Third World is of prime importance to all but will not be reached with good will alone. Since self-interest is a more common attribute than philanthropy, the most promising way forward is for the richer countries to realise that investment leading to Third World economic growth will result in billions of new consumers.

5. The educational maturity of a country should not be measured by the number of its universities but by how it teaches its younger, poorer and simpler citizens.

   Asia Survey in The Economist, October 30, (1989) comparing percentage GDP spent on university education in Venezuela and South Korea.

6. It is unfortunate that the most important issues in society that are organised by the state, such as education and the environment, require policies that outlast the lifetime of the term of any democratically elected government.

7. Problems seem to become easier to solve the further one is separated from them. This holds true for all kinds of problems, from simple domestic nuisances to major political confrontations. The emphasis is on the "seem".

8. Nationalism, in its milder form, is harmless enough when confined to sports but is dangerous in its extreme version when it impinges on international relations. A practical way to reduce extreme nationalism without affecting its milder form is to make sure that history is taught in primary and secondary education by a mixture of native and foreign teachers. The idea that one's country can do no wrong and has won all battles would, hopefully, not take hold in young, impressionable minds.

9. Eating sandwiches with the aid of a knife and fork is the first clear sign to a newcomer in the Netherlands of the prodigious efficiency of the Dutch.

10. The phonetic nature of the Dutch language is being changed by the continual use of English words. This can be seen from the following considerations: plan or plen; planning or plenning, ham or hem, Big Mac or Bik Mek, flat or flet, snapt u het or snept u het ?

Letter in Volkskrant, 11 December 1993.