Chapter 1

Introduction

1.1 The Semiconductor Industry

The semiconductor industry is one of the world's largest grossing market sectors, with an annual turnover of hundreds of billions of dollars. At the heart of this behemoth lies the simple metal-oxide-semiconductor field effect transistor (MOSFET), which is the most abundant man-made structure ever devised. A microprocessor chip in a standard home computer typically contains tens of millions of these transistors at the time of writing.(1)

The combination of silicon and an excellent insulator in the form of its oxide, SiO_2, forms the basis for approximately 95% of the semiconductor industry. This dominance arises for a number of reasons. Firstly, silicon is easily obtainable (comprising a quarter of the Earth's crust) and therefore inexpensive. Secondly, the oxide layer may be manufactured to a very high standard, with few traps either within it, or at the interface with the semiconductor. Finally, the chemical and mechanical properties of the oxide may be exploited during device fabrication for masking and structuring steps.(2)

The continual growth of the semiconductor industry is founded on a forty-year-old law, first stated by Moore.(3) In 1965 he realised that the density of components on an integrated circuit doubled every eighteen months, and this has been proven accurate since that time. In 1992 the first *Technology Roadmap for Semiconductors* was

published. Many more have followed, which outline all the relevant dimensions and performance targets that must be met in order to continue the phenomenal progress.(4)

This progress has arisen solely as a result of steadily improving processing techniques, which have enabled all of the device dimensions to be reduced. Perhaps the most important of these has been the enhancement in lithography techniques, which have allowed ever more intricate masks to be used. However, it is becoming increasingly apparent that the improvement of standard silicon complimentary metal-oxide-semiconductor (CMOS) technology cannot continue forever in its present form and that a "roadblock" is likely to be reached over the course of the next decade.

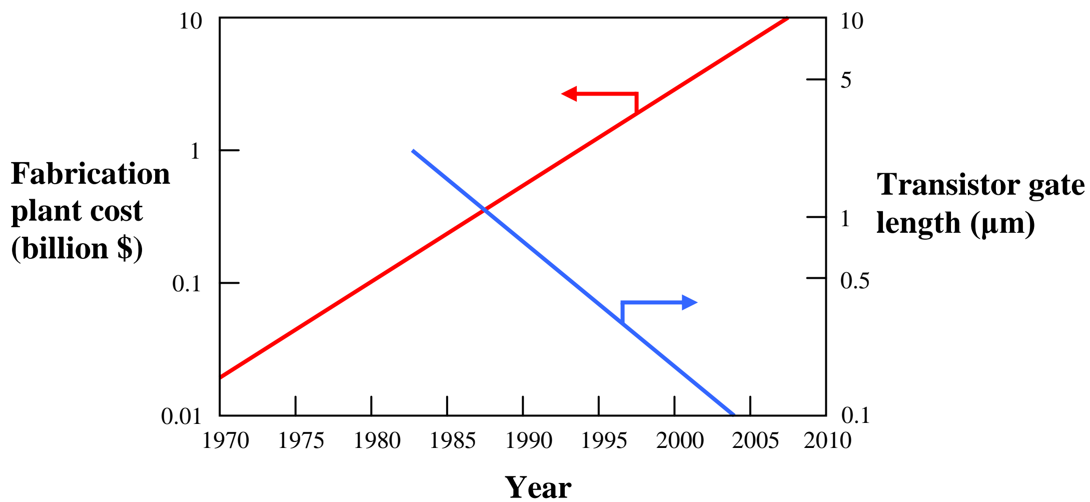


Figure 1.1. The reduction in transistor gate length and the increasing cost of production lines as a function of time (after Paul).(5)

The cost of new fabrication plants is fast approaching the point at which no individual company would be able to bear the financial burden of commissioning one (figure 1.1).(5) Even worse, the dimensions of MOSFETs are approaching a scale at which they will be composed of just a few hundred atoms and quantum effects such as

tunnelling (which are already problematic) will prevent further improvements in transistor performance due to elevated leakage currents. Additionally, random dopant induced parameter fluctuations will result in a great variation in the properties of devices across a wafer.[6]

However, all is not necessarily doom and gloom. There exist many materials that possess far superior electrical properties to silicon; indeed some of these have already found employment in niche markets, such as optical devices or the high frequency components of mobile communication devices. These materials have not been able to move into the mainstream simply because of economic considerations, and because they find it hard to compete with the vast accumulation of knowledge in the silicon industry. One particularly promising material, which has yet to be exploited to its fullest potential, is the element germanium.

1.2 Germanium

It is somewhat ironic that the first transistor ever created (in 1947) used germanium as an active material and after half a century of neglect, the element is again beginning to find favour with the semiconductor industry. Despite having lower electron and hole mobilities, silicon rapidly displaced germanium as the material of choice due to its natural abundance and useful natural oxide.

Combining germanium with silicon to form an alloy of the two is a promising way of exploiting the benefits that it can offer (see chapter 2 for a description of these), without overcomplicating the issue of processing. Because both silicon and germanium are from group IV in the periodic table and both assume the diamond

structure at room temperature and atmospheric pressure, their alloy (SiGe) also has the same crystallographic and electronic properties.[7] This allows it to be incorporated into a standard silicon MOSFET fabrication process without too many alterations and most importantly from the point of view of the industry, without greatly increasing costs.

SiGe is in fact already in use in bipolar junction transistors (BJTs) as an effective way of increasing the current gain or, by sacrificing this, the operating frequency of the device.[8] In this application, it is the smaller bandgap of SiGe compared to silicon that is exploited, as the barrier to electron transport between the n-type emitter and collector contacts (created by the inclusion of a p-type base region) is reduced.[9] Additionally, the inclusion of germanium increases the solid solubility of boron in the base region, allowing higher doping concentrations and a lower base resistance as well as hugely suppressing out-diffusion of boron. Such SiGe HBTs (heterojunction bipolar transistors) are in commercial production.

Whilst SiGe has therefore already proven its worth, incorporating it into MOSFETs is proving more problematic, not least because the very feature which makes it so attractive for use in HBTs, its smaller bandgap, can create an unwanted junction leakage current. Instead of attempting to replace the silicon channel, some experimental devices have used SiGe as a polycrystalline gate electrode, or in the source and drain region to exploit the increased solid solubility for boron.[10] This allows for tighter doping profiles or a lower resistance by virtue of an increase in the maximum dopant concentration.

Importantly though, SiGe also has the potential to greatly improve the carrier mobility in an MOSFET, allowing for higher operating frequencies and permitting some breathing space from the relentless dimension down-scaling. The simplest way to introduce SiGe to the production line is by depositing only a thin layer (~ 10nm) directly onto a silicon substrate, with a further thin silicon cap layer on top of this for cleaning and oxidation purposes. The lattice constant of germanium is 4.2% larger than that of silicon and an alloy of the two has been found to approximately follow Vegard's law.(7) The thin SiGe layer assumes the lattice constant of silicon and becomes compressively strained, which has important implications for its transport properties. The mismatch between the energy bands of silicon and strained SiGe mostly occurs in the valence band, giving rise to a quantum well for holes, in which their mobility can be much higher than in silicon or indeed bulk germanium.(11) The structure is suitable for a relatively standard CMOS process because the vast majority of the wafer is still silicon, most importantly the upper surface. At the time of writing it looks unlikely that these fully pseudomorphic devices will ever enter mainstream production because of inherent problems with employing a buried channel for carriers, although it is possible that new dielectric materials will circumvent this difficulty.

Fortunately SiGe also offers the possibility of creating surface channel devices with improved electrical properties. If a sufficiently thick layer is grown onto a silicon substrate then rather than take the lattice constant of the layer beneath it, it will relax and assume its own natural lattice constant. If a thin layer of material is then grown on top of this SiGe *virtual substrate* then it assumes the lattice constant of the SiGe layer directly beneath it, rather than the underlying silicon substrate. This allows a

heterostructure to be created with biaxial tensile strained silicon as its uppermost layer. Strained silicon has the potential to greatly increase the mobilities of both electrons and holes, whilst retaining nearly all of the advantages that have allowed silicon to occupy such a dominant position in the semiconductor industry.[12]

1.3 Current Investigation

This thesis is an investigation into the properties of strained silicon devices. Whilst there has already been substantial investment by industry giants in this area, there are still several aspects of the Si/SiGe heterostructure that are poorly understood. In particular, there has been far less research into pMOSFET devices (which use holes as charge carriers) than nMOSFETs (which use electrons), largely because a higher degree of strain is required for noticeable hole mobility enhancements and this has severe implications for growth and device processing. Consequently, properties such as the valence band offset between strained silicon and relaxed SiGe are still not well known and the hole mobility enhancements predicted by theory have not been realised experimentally. It is the aim of this investigation to clarify some of the remaining grey areas concerning strained silicon technologies, with a particular focus on pMOSFET devices, which have been neglected so far.

Chapters 2 and 3 are concerned with the relevant background material to this investigation, the latter chapter dealing with the necessary characterisation techniques. Chapter 4 presents original work on self-heating in strained silicon MOSFETs, with a consideration of the requirements that this phenomenon places on the virtual substrate. Chapter 5 presents the results of investigations into three batches of strained silicon wafers. Chapter 6 deals with the simulation of one of these batches,

subsequent to characterisation, from which the valence band discontinuity between tensile strained silicon and relaxed SiGe is extracted.

References

1. D. Boggs *et al.*, "The Microarchitecture of the Intel® Pentium® 4 Processor on 90nm Technology." *Intel Technology Journal.*, vol. 8, no. 1, 2004. Website: http://developer.intel.com/technology/itj/2004/volume08issue01/
2. S. M. Sze, "VLSI Technology", 2nd edition, McGraw-Hill Book Co., pp. 98, 1988.
3. G. Moore, "Cramming More Components onto Integrated Circuits", *Electronics*, vol. 38, no. 8, 1965.
4. International Technology Roadmap for Semiconductors. Website: http://public.itrs.net/
5. D. J. Paul, "Silicon-Germanium Strained Layer Materials in Microelectronics", *Adv. Mater.*, vol. 11, no. 3, pp. 191-204, 1999.
6. A. Asenov, A. R. Brown, J. H. Davies, S. Kaya and G. Slavcheva, "Simulation of Intrinsic Parameter Fluctuations in Decananometer and Nanometer-Scale MOSFETs", *IEEE Trans. Elec. Dev.*, vol. 50, no. 9, pp. 1837-1852, 2003.
7. F. Schäffler, "High-Mobility Si and Ge Structures", *Semicond. Sci. Technol.*, vol. 12, pp. 1515-1549, 1997.
8. Y. H. Xie, "SiGe Field Effect Transistors", *Materials Science and Engineering*, vol. 25, pp. 89-121, 1999.
9. Y. Taur and T. H. Ning, "Fundamentals of Modern VLSI Devices", Cambridge University Press, pp. 363-368, 1998.
10. C. H. Chen et al., "Improved Current Drivability and Poly-Gate Depletion of Submicron PMOSFET with Poly-SiGe Gate and Ultra-thin Nitride Gate Dielectric", *Solid-State Electronics*, vol. 46, no. 4, pp. 597-599, 2002.
11. T. E. Whall and E. H. C. Parker, "SiGe-Heterostructures for CMOS Technology", *Thin Solid Films*, vol. 367, pp. 250-259, 2000.
12. C. K. Maiti, N. B. Chakrabarti and S. K. Ray, "Strained Silicon Heterostructures: Materials and Devices", Institution of Electrical Engineers, pp. 307-318, 2001.

Chapter 2

Background to Strained Silicon MOSFETs

2.1 Introduction

This chapter serves as a summary of the important issues relating to biaxial tensile strained silicon MOSFETs. Basic MOSFET operation is outlined and the scattering mechanisms encountered in a standard silicon device are discussed, with attention paid to the conditions under which each mechanism becomes dominant. The band structure of a strained silicon device is then introduced, together with an explanation of the effect of tensile strain on the electrical properties of silicon and how an improvement over conventional CMOS technology may be realised.

Intel® has recently introduced uniaxial strained silicon at the 90nm technology node, at which some transistors have a written gate length of less than 60nm. If biaxial tensile strained silicon ever enters mainstream production, it will be at similar very short channel lengths. This chapter therefore also considers the short channel effects encountered in MOSFETs and the likely effect of these on strained silicon. Finally the field of virtual substrates is reviewed, since a high quality virtual substrate is of vital importance to the integration of SiGe into production. The problem of excessive self-heating introduced by the use of virtual substrates is summarised.

2.2 Basic MOSFET Operation

Figure 2.1 is a schematic diagram of a simple pMOSFET. The gate, oxide and bulk silicon act as a parallel plate capacitor with the oxide layer as the dielectric. If the gate is negatively biased with respect to the source and substrate, the majority carriers (electrons) in the bulk n-type silicon are forced away from the Si/SiO_2 interface and a depletion region is formed. If the bias is increased further, the surface becomes attractive to holes, which are in plentiful supply in the p^+ doped source and drain regions, and an inversion layer of holes is formed in the normally n-type silicon.(1) Holes migrate across this channel from source to drain, as the drain is also negatively biased with respect to the source. Increasing the gate bias increases the concentration of holes at the semiconductor surface, and allows more current to flow. The gate can therefore be used to modulate the current flow between source and drain, giving rise to the switching operation in integrated circuits.

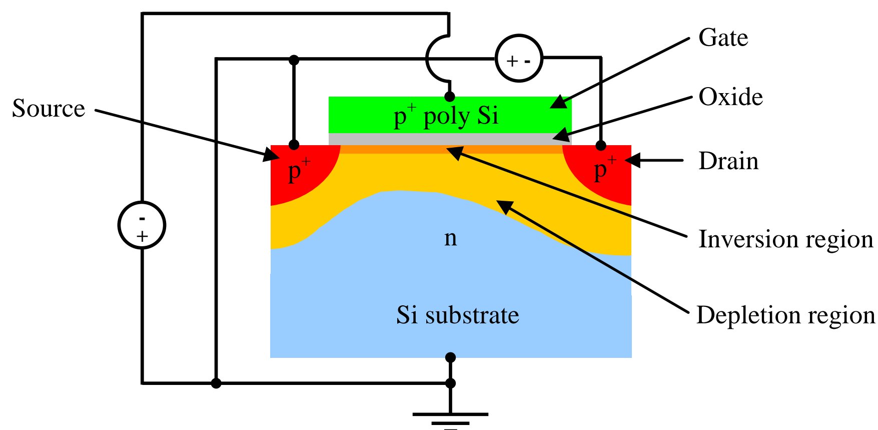


Figure 2.1. Diagram of a simple pMOSFET, showing the bias applied to each of the four contacts during basic operation.

To simplify the description of effects within a MOSFET, the *gradual channel approximation* is commonly made.[(2)] This assumes that the horizontal electric field strength along the channel of the MOSFET is much smaller than the vertical electric field and enables the use of the one-dimensional form of Poisson's equation. Under the *charge-sheet approximation*, it is further assumed that all the inversion charges are located at the semiconductor surface in a sheet, and there is no potential drop or band bending across the inversion layer. Thus it can be written:

$$Q_i(x) = C_{ox}(V_{gs} - V_t - V_c(x)), \tag{2.1}$$

where Q_i is the charge in the inversion layer per unit area, C_{ox} is the gate oxide capacitance per unit area, V_{gs} is the gate voltage relative to the source contact, V_t is the threshold voltage (see section 3.5.2) and V_c is the channel potential with respect to the source. x is the distance along the channel, being zero at the source end and equal to L, the channel length, at the drain end. The carrier velocity along the channel, v, is related to the horizontal electric field by the carrier mobility, μ:

$$v = \mu \frac{dV_c}{dx}. \tag{2.2}$$

The drain current, I_{ds}, which in this simple case has the same value at any point along the channel, may now be expressed:

$$I_{ds} = C_{ox} \mu W \frac{dV_c}{dx}(V_{gs} - V_t - V_c(x)), \tag{2.3}$$

where W is the channel width. By integrating from source ($x = 0$, $V_c(0) = 0$) to drain ($x = L$, $V_c(L) = V_{ds}$), the final expression for I_{ds} is obtained:

$$I_{ds} = \mu \frac{W}{L} C_{ox} \left[(V_{gs} - V_t)V_{ds} - \frac{V_{ds}^2}{2} \right], \tag{2.4}$$

where V_{ds} is the potential of the drain with respect to the source.

The channel current therefore has a maximum at $V_{ds} = V_{gs} - V_t$ and the MOSFET enters the *saturation* region. This may be understood by considering the fact that as V_{ds} is increased, the current increases but the inversion charge density at the drain decreases (equation 2.1), until it finally goes to zero and the inversion region at the drain end of the channel vanishes at the onset of saturation. This is known as *pinch-off* and a further increase in V_{ds} results in the pinch-off point being moved towards the source as the extra potential is dropped across the high resistance region near the drain. The current is maintained at the same magnitude. The saturation current, I_{dsat}, is therefore given by:

$$I_{dsat} = \mu \frac{W}{2L} C_{ox} \left(V_{gs} - V_t\right)^2. \qquad (2.5)$$

As gate length is reduced to submicron levels, short channel effects (SCEs) begin to affect device operation and the gradual channel approximation becomes invalid. SCEs include barrier lowering, punchthrough, velocity saturation, hot carriers, quantum mechanical tunnelling and velocity overshoot.(3) These effects are discussed in section 2.5.

2.3 Carrier Mobility and Scattering

In section 2.2, the carrier mobility was introduced as a quantity relating the carrier velocity to the applied electric field. Increasing carrier mobility will therefore result in an increase in MOSFET drain current when it is in its on-state, which is an important parameter in terms of CMOS circuit performance. The mobility of electrons in undoped bulk silicon at 300K is approximately $1350 cm^2V^{-1}s^{-1}$, whilst the mobility of holes is a comparatively modest $450 cm^2V^{-1}s^{-1}$,(4) explaining the

requirement for pMOSFETs to be made wider than nMOSFETs in CMOS circuits, in order to match the current drive of the two types of transistor. Unfortunately, several effects encountered in MOSFETs have a detrimental effect on mobility, by scattering carriers.

In order to consider the effects of a particular scattering mechanism on carrier transport, carriers are frequently treated as Bloch waves.(5) A Bloch wave is mathematically defined by:

$$\psi_k = u_k e^{ikz}, \tag{2.6}$$

where the eigenfunction $u_k(z) = u_k(z + a)$, reflecting the periodicity of the crystal lattice (a is the lattice constant). Bloch waves move through the lattice unimpeded by the crystal potential. However, occasionally a carrier encounters a perturbation due to a lattice vibration or an impurity/defect. It is scattered and a carrier with crystal momentum **p** undergoes a transition to a state **p'**.

A common feature of scattering models is that a scattering potential is chosen such that the scattering rate from **p** to **p'** is proportional to the density of states. The more states that are vacant at a given energy, the higher the probability of a carrier scattering into one of those states, and the higher the scattering rate.

2.3.1 Phonon Scattering

The displacement of atoms by lattice vibrations affects the local band structure and changes the conduction and valence band energies. This acts as a perturbing potential for Bloch waves that encounter these vibrations, or *phonons*. There are two branches of phonons: optical phonons and acoustic phonons.(6) The difference between the two

types is depicted in figure 2.2. For optical phonons, adjacent atoms move in opposite directions whereas for acoustic phonons, adjacent atoms in each half-loop move in the same direction.

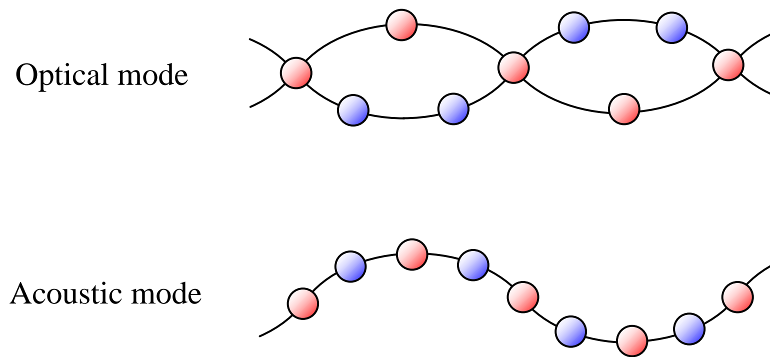


Figure 2.2. Diagram depicting the difference between optical and acoustic phonons having the same wavelength.

Due to the greater distortion between adjacent atoms when their displacements are in opposite directions, optical phonons have a much higher minimum energy than acoustic phonons. Acoustic phonons have negligible energy compared to carriers, so scattering from them is effectively elastic.

The degree of phonon scattering depends on the number of active phonon modes present, which is a strong function of temperature. Phonon scattering is frequently the limiting mechanism to carrier mobility at room temperature. One of the key aims of employing a strained silicon layer, as the channel of a device, is to greatly reduce intervalley phonon-assisted carrier scattering (section 2.4). Carriers may emit as well as absorb phonons. Because optical phonons have higher energy, there is a threshold for optical phonon emission, above which the optical phonon scattering rate increases

markedly. This is the mechanism that causes velocity saturation, which is discussed in section 2.5.

2.3.2 Surface Roughness

When confining an inversion layer against a junction between two materials, the roughness of the interface will scatter carriers. This is the case in an MOSFET, where a thin "sheet" of carriers, known as a two dimensional electron/hole gas (2DEG/2DHG) is confined against the silicon/oxide interface. Typically, the form of the interface roughness is assumed to be distributed about the characteristic length, Λ, in a Gaussian peak of amplitude Δ. Interface roughness scattering will be at a maximum when its characteristic length is of the order of nanometres, since this is the typical wavelength of carriers in the channel.(7) This is termed *micro-roughness*.

Scattering due to interface roughness is an example of scattering at a static potential. Carriers are able to rearrange themselves in response to this potential and screen it in such a way that its influence is greatly reduced. However, a high carrier sheet density will usually only result from the application of a large vertical electric field, which has the effect of forcing the carriers against the interface and making the problem of interface roughness scattering worse. This effect overcomes any benefit gained from screening, and scattering due to interface roughness is at its greatest when the vertical field is large.(8) Interface roughness scattering is of great importance for modern MOSFETs, which operate with effective vertical fields in excess of $1MVcm^{-1}$.

2.3.3 Ionised Impurities

Scattering at ionised impurities is a further example of a static scattering mechanism. Ionised impurities in the channel region are likely to result from doping steps during fabrication. For example, ion implantation steps have an associated straggle, meaning that if the desired dopant depth is a few tens of nanometres from the surface (as is the case for a punchthrough stop), there will be a distribution of dopant atoms back to the surface. Annealing stages during device fabrication can exacerbate the problem if the temperature is high enough and the duration long enough (this is termed the *thermal budget*) to allow diffusion of dopant atoms to become an issue. The semiconductor industry tends to avoid this problem by the use of rapid thermal anneals (RTAs), whereby the temperature of the wafer is raised to 1000ºC or more for just a few seconds.[9]

Carriers with a high energy are not scattered by fixed charges to the same degree as low energy carriers. In addition, because ionised impurities may be screened, their effect is greatest when the number of carriers in the channel is low. This scattering mechanism is therefore usually only limiting in weak inversion.

2.3.4 Alloy Scattering

The formation of a SiGe alloy introduces a further scattering mechanism.[10] The germanium atoms disrupt the silicon crystal and alter the band structure locally, in addition to modifying it for the entire crystal. This local disruption of the band structure results in short range scattering potentials, however the mechanism is not well understood. As one would expect, it appears that the scattering rate is at a maximum when the germanium content of the alloy is 50%, but it is unclear exactly

how great the effect of alloy scattering is on carrier mobility and whether or not it is a screened mechanism. Alloy scattering is not expected to be a problem for strained silicon devices, in which it is hoped that the vast majority of carriers will be confined to the pure silicon channel. However, for pMOSFETs in particular there may be an appreciable proportion of carriers in the relaxed SiGe underlying the channel, so this mechanism is mentioned for completeness.

2.4 Strained Silicon

2.4.1 Band Structure

The fundamental band gap in both silicon and germanium is indirect and remains so for all compositions of the $Si_{1-x}Ge_x$ alloys.(11) The conduction band minima in silicon occur along the <100> crystal directions (Δ minima) and are six-fold degenerate. However, the conduction band minima in germanium are located in the <111> directions (L minima) and are eight-fold degenerate. The band structure of SiGe is silicon-like for all compositions up to approximately 85% germanium,(12) when there is a transition to germanium-like behaviour. The bandgap of unstrained $Si_{1-x}Ge_x$ therefore decreases monotonically from 1.12eV to 0.66eV (at 300K) as germanium content is increased from 0 to 100%, with a kink at 85% (figure 2.3).

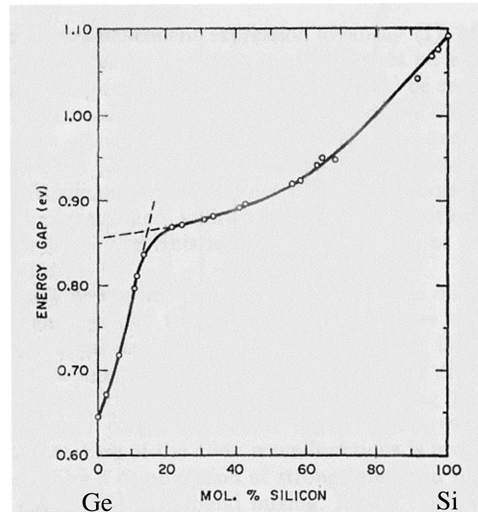


Figure 2.3. Variation in the energy gap of SiGe as a function of composition (after Braunstein *et al.*).[13]

As has been mentioned in section 1.2, the use of SiGe allows the creation of a thin layer of strained silicon under tetragonal (biaxial tensile) strain. This strain has large implications for the band structure of the semiconductor. Figure 2.4(a) shows the constant energy ellipsoids of the six valleys of the unstrained silicon conduction band. Figure 2.4(b) shows the effect of tetragonal strain on these valleys. Tetragonal strain has the effect of lifting the six-fold degeneracy of the conduction band in silicon into a two-fold and four-fold degenerate set.[14] The deformation potential of the strain lowers the energy of the two valleys with their long axis perpendicular to the Si/SiGe interface. The amount of energy lowering is dependent on the degree of strain. It has been theoretically predicted [15] that approximately 0.8% strain, resulting from a $Si_{0.8}Ge_{0.2}$ alloy, provides sufficient lowering that only the two-fold degenerate valleys are occupied at room temperature and low effective vertical field.

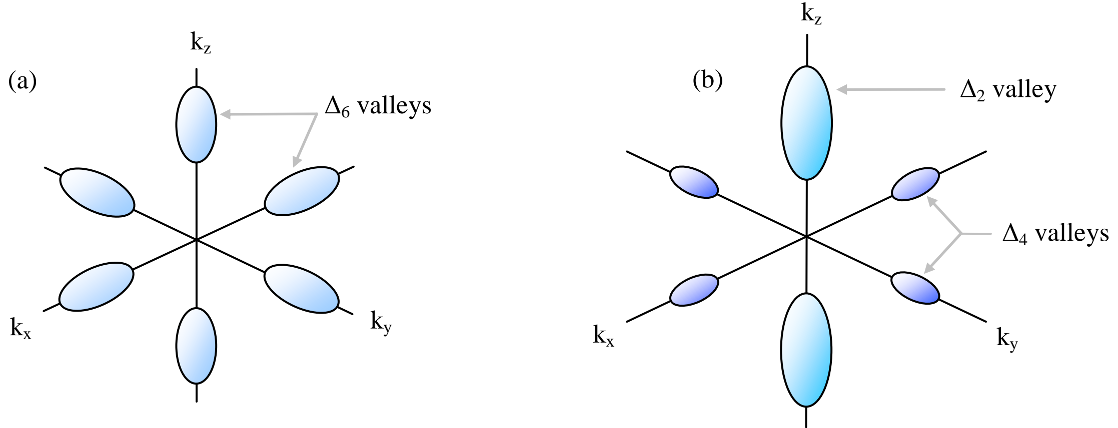


Figure 2.4. (a) Six-fold degenerate energy ellipsoids of the conduction band of unstrained silicon. (b) The same ellipsoids in the presence of tensile strain.

Intervalley carrier scattering may only occur between degenerate minima [16] and is assisted by phonons. The addition of tensile strain to silicon therefore has the effect of disallowing much of the intervalley scattering that previously occurred because there are now far fewer possible final states for a carrier to scatter into. In the case that only the two-fold degenerate valleys are populated, the only intervalley scattering that occurs is between them.

Strain has a similar effect on the valence band of silicon. Figure 2.5 shows the arrangement in the unstrained case. The light and heavy hole bands are degenerate at the zone centre. The labels of light and heavy hole bands arise because of the effective masses of carriers occupying these states. It is relatively simple to show that:

$$\frac{1}{\hbar^2}\frac{d^2E}{dk^2}=\frac{1}{m^*}. \tag{2.7}$$

Here m^* is the effective carrier mass, which is different from the electron rest mass due to interactions with the lattice potential.[17] Hence the light hole band has greater

curvature than the heavy hole band. The third band in figure 2.5 is the split-off band, which is 44meV higher in energy due to spin-orbit interactions (for holes, energy is measured in the opposite sense to electron energy).

The addition of strain lifts the degeneracy of the light and heavy hole bands and causes the split-off band to be further removed from the other two, reducing further the possibility of any interband scattering events involving it. In the case of tensile strain, the light hole band is reduced in energy relative to the heavy hole band and mass inversion occurs, such that holes in the light hole band are actually heavier than those in the heavy hole band. The effect of strain on the valence band of silicon is quite complex and has severe implications for many models that assume the energy bands have parabolic energy-momentum relations and spherical equal energy surfaces.[18] The advantage of this assumption is that a constant effective mass may be assigned.

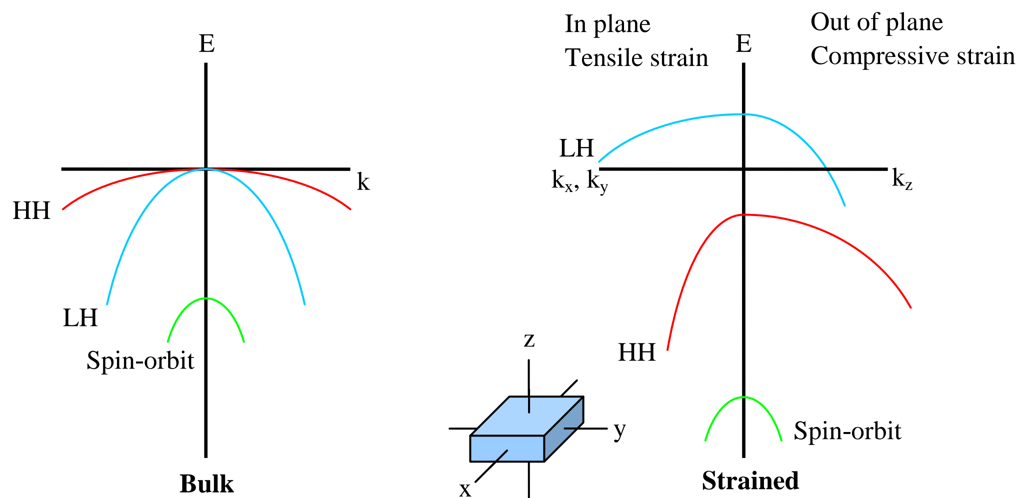


Figure 2.5. The effect of strain on the valence bands of silicon (after Rim *et al.*).[20]

It has been shown experimentally that a greater degree of strain is required to cease carrier scattering between the light and heavy hole bands than is required to confine all the conduction electrons to the two-fold degenerate minima.[19] To completely eliminate interband scattering in a strained silicon overlayer may require SiGe buffers with a germanium composition as high as 40%. Rim *et al.* [20] give the energy splitting between the light and heavy hole valence bands as 38meV per 10% germanium in the buffer layer. In comparison, the splitting between the Δ_2 and Δ_4 levels in the conduction band is thought to be 67meV per 10% germanium.

The carrier effective mass may be reduced under the influence of strain. Mobility is inversely related to the effective mass by the equation:

$$\mu = \frac{q\tau}{m^*}, \qquad (2.8)$$

where τ is the relaxation time and q is the electronic charge. Because the carrier effective mass and the phonon scattering rate are both dependent on the atomic lattice, carrier mobility is dependent on the crystal orientation. For holes, the highest mobilities are achieved on wafers with an (011) surface orientation rather than the usual (001). However, since electrons suffer reduced mobility in this orientation, IBM® has recently created a hybrid-orientation technology in which CMOS is fabricated on a hybrid substrate with different crystal orientations to achieve a significant pMOSFET performance enhancement.[21]

The combination of a reduction in m^* and an increase in τ, brought about by the reduction of interband scattering, is thought to yield the increase in mobility seen when silicon is strained. Figure 2.6 shows the low field mobility enhancements offered by strained silicon as a function of strain, as predicted by Oberhüber *et al.*[22]

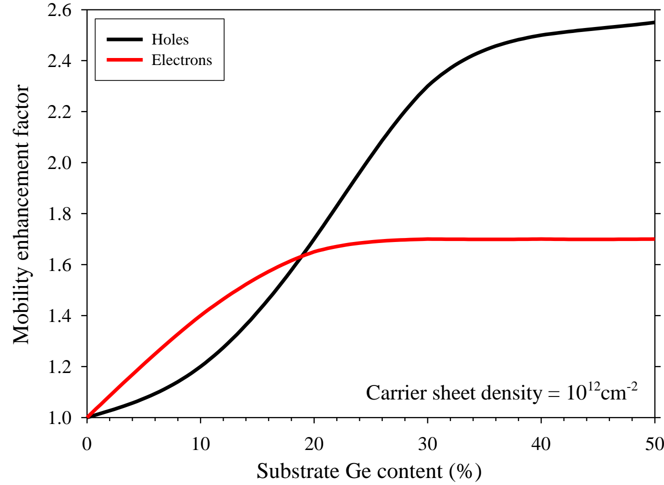


Figure 2.6. The theoretically predicted low field mobility enhancement offered by strained silicon at 300K (after Oberhüber *et al.*).[22]

However, Fischetti *et al.* [23] argue that the removal of the six-fold degeneracy in the conduction band of silicon and the reduction of effective mass brought about by tensile strain is insufficient to explain the electron mobility enhancement seen for high carrier sheet densities. The electron confinement caused by the inversion potential at the silicon/oxide interface also has the effect of lifting the degeneracy. The band bending necessary to create an inversion carrier sheet density of 10^{13}cm^{-2} in relaxed silicon should be sufficient to confine ~ 75% of electrons to the Δ_2 minima, whereas 1% strain will confine ~ 90% of electrons. We would therefore expect the electron mobility enhancement offered by strained silicon to diminish substantially at large vertical effective fields. As this is not the case, it appears there may be some other mechanism at work. By fitting a model for interface roughness limited mobility to existing experimental data, Fischetti *et al.* are able to demonstrate that the mobility enhancement seen in strained silicon under high vertical fields may well be due to reduced roughness at the semiconductor/oxide interface. However, it remains unclear

how a small change of lattice constant can affect the properties of the oxidation process and the smoothness of the interface to the degree required (up to a 50% reduction is required to explain some experimental results).

2.4.2 Band Alignment

Strained silicon on a relaxed SiGe buffer layer has the type II band alignment,(24) meaning that the conduction and valence band discontinuities occur in the same direction, as shown in figure 2.7. The conduction band offset is helpful in confining electrons to the strained silicon layer but unfortunately the valence band offset works against the strained silicon pMOSFET designer as it may be energetically favourable for holes to occupy the SiGe buffer under low negative gate bias conditions. The magnitude of the band discontinuities as a function of germanium percentage is not well known, although it is generally agreed that the conduction band offset is somewhat larger than that of the valence band. For example, Maiti *et al.* (25) give the valence band offset between strained silicon and $Si_{0.7}Ge_{0.3}$ as 180meV whilst Rieger and Vogl (26) predict that it is as little as 70meV. It is important to know these quantities accurately, as they have a large impact on the threshold voltage of a strained silicon device. It is one of the aims of the simulations performed in chapter 6 to determine this offset more accurately.

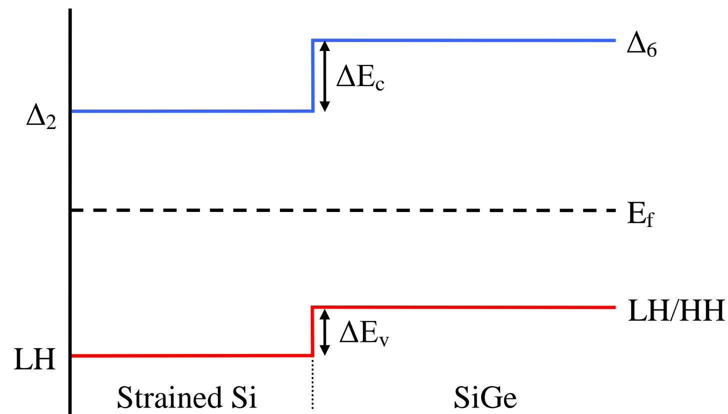


Figure 2.7. The type II band alignment between strained silicon and relaxed SiGe.

2.4.3 Critical Thickness

The problems that can be encountered with parasitic conduction in pMOSFETs fabricated on strained silicon can be overcome if the strained silicon channel is sufficiently thick, so that the Si/SiGe interface is greatly removed from the Si/SiO_2 interface. Under these conditions, negligible band bending will occur in the region of the SiGe when a gate bias is applied and almost all of the conduction holes will be confined to the high mobility strained silicon layer. Unfortunately there is an upper limit on the thickness of a strained silicon layer before it becomes energetically favourable for it to relax with the introduction of interfacial misfit dislocations. This limit is known as the *critical thickness* and varies with strain. Figure 2.8 is taken from the work of Samavedam *et al.*,[(27)] which is based on original calculations made by Matthews and Blakeslee [(28)] and the work of Bean *et al.*.[(29)] The channel thicknesses of some existing devices are also included for comparison.

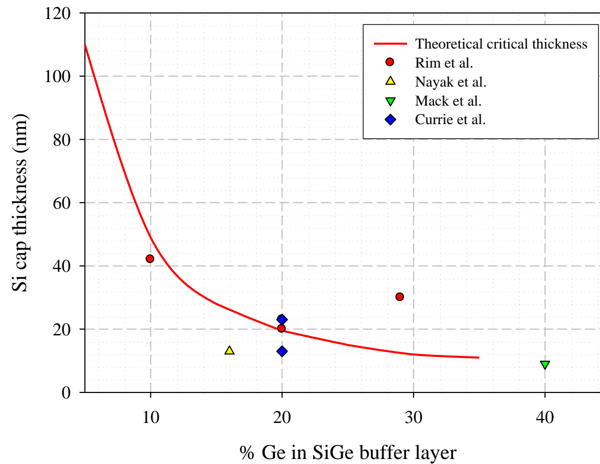


Figure 2.8. The theoretical critical thickness for strained silicon on an unstrained SiGe alloy layer (after Samavedam *et al.*),[27] together with the specification of some existing research devices.

It can be seen from the work of Rim *et al.* [30] on a $Si_{0.71}Ge_{0.29}$ buffer that it is possible to exceed the critical thickness and still retain a strained layer. In fact there is also a metastable critical thickness at low growth temperatures (below about 550°C). In this regime, the lowest energy state would be for the strained silicon layer to relax but there is insufficient thermal energy to create the dislocations required to do so, and the layer remains strained. As yet the metastable critical thickness of strained silicon as a function of growth temperature has not been quantified. Extra care must be taken when processing such wafers, since high thermal budgets can easily cause relaxation of the silicon.

Taking into consideration the fact that cleaning and oxidation stages are likely to consume at least 2 - 3nm of strained silicon,[31] it is clear that the creation of strained silicon MOSFETs on high germanium content virtual substrates can be extremely problematic and explains why much of the research in this field has been focused on

strain levels of 1% or less. However, to unlock the largest hole mobility enhancements is thought to require up to 2% strain, corresponding to a germanium content in the virtual substrate of over 40%.

2.4.4 Design Issues

This section deals with some of the important criteria to consider when designing strained silicon pMOSFETs. From this perspective pMOSFETs are generally more complex than their nMOSFET counterparts, owing to the need to employ higher germanium concentrations to realise an appreciable mobility enhancement, and the problems of parasitic conduction in the SiGe layer. It should be noted that from a fabrication perspective, strained silicon nMOSFETs can also be problematic because of the elevated diffusion of n-type dopants in SiGe.

When dealing with high germanium content virtual substrates, it is of primary importance to use the thickest strained silicon layer possible in order to minimise parasitic conduction, as mentioned previously. Another possibility, introduced by Rim *et al.*,[30] is to create a graded region between the SiGe and the silicon known as a *gradeback layer*. Instead of growing pure silicon directly onto the buffer layer, the germanium content is linearly graded to zero over a region of a few nanometres. This has the effect of smoothing out the sharp valence band offset between the two regions and can help reduce the proportion of holes that are confined outside the silicon region (figure 2.9), even though the thickness of pure silicon must be somewhat reduced in order to keep the layer thermally stable.

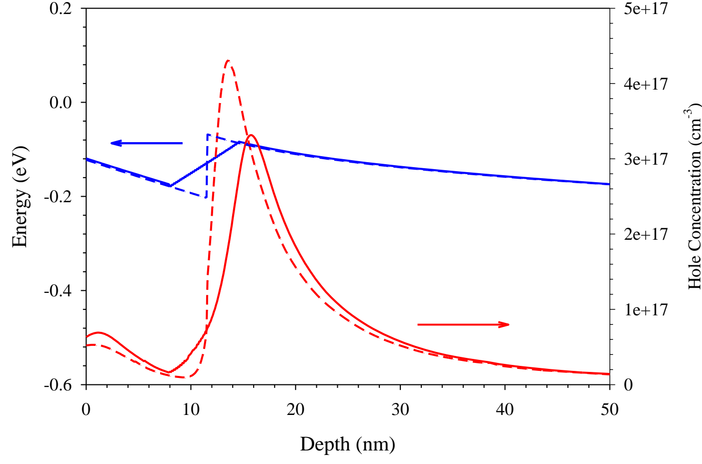


Figure 2.9. Effect of adding a gradeback layer (solid lines) on the valence band and vertical hole profile of a strained silicon device.

For small devices with a thin oxide, the importance of using a thick strained silicon layer is reduced. This is because a thin oxide leads to high vertical fields and a larger degree of band bending at the surface than at the Si/SiGe interface. This is evident from a consideration of the gate to silicon and gate to SiGe capacitances in the two cases of thick and thin gate oxides. Ignoring the influence of interface traps, the gate to channel capacitance is simply given by C_{ox}. The gate to SiGe capacitance per unit area, C_{SiGe}, is given by:

$$\frac{1}{C_{SiGe}} = \frac{1}{C_{ox}} + \frac{1}{C_{SS}}, \tag{2.9}$$

where C_{SS} is the areal capacitance of the strained silicon layer. For a 10nm thick silicon channel, C_{SS} is approximately equal to 10^{-2}Fm^{-2}. A 100nm thick gate oxide has a capacitance of approximately 4×10^{-4}Fm^{-2} whereas a 1nm oxide has a capacitance of 4×10^{-2}Fm^{-2}. It is therefore apparent that for the thick oxide case, C_{ox} is much smaller than C_{SS}, and C_{SiGe} becomes equal to C_{ox}. However, when a thin oxide is employed, C_{ox} is approximately equal to C_{SS}, and C_{SiGe} becomes close to $C_{ox}/2$. The capacitive coupling between the gate and channel relative to the coupling between the

gate and SiGe is therefore greater when a thin oxide is employed and parasitic conduction in the SiGe layer is significantly reduced.

A further method of reducing parasitic conduction is the use of doping in the SiGe buffer layer underneath the channel. If additional n-type doping is added to this region then significant band bending can occur and hole population is greatly reduced (figure 2.10). Parasitic conduction is unlikely to be completely eliminated by this method since the valence band offset between the two layers still remains, but it can be highly effective. Control of parasitic conduction in this way occurs in state of the art devices as a matter of course because similar retrograde doping profiles are typically employed to control the threshold voltage and reduce undesirable short channel effects. Combined with a thin oxide, this means that parasitic conduction should not be an issue in a production device. It is, however, of major concern in some of the research devices encountered in this work.

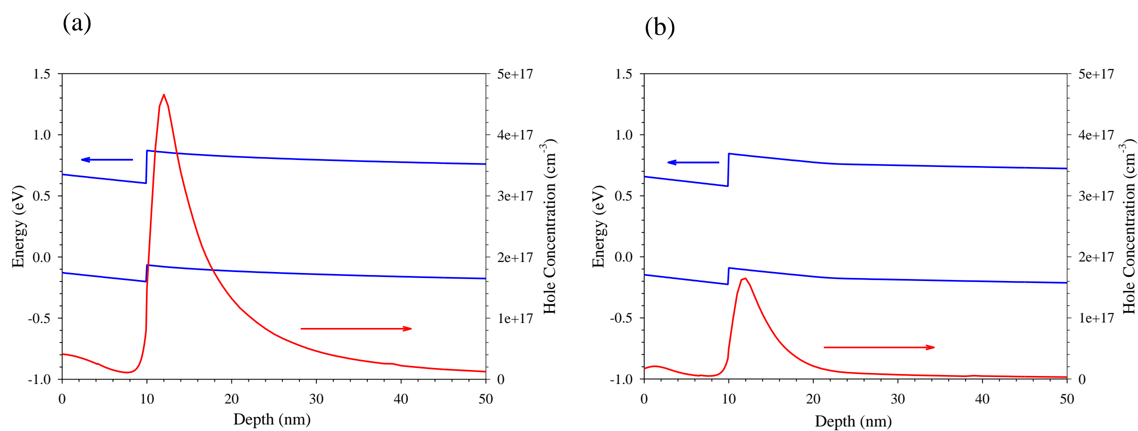


Figure 2.10. Band bending created by the introduction of n-type doping to the SiGe buffer below a strained silicon channel. (a) $N_d = 1\times10^{15}$cm^{-3}, (b) $N_d = 8\times10^{17}$cm^{-3}.

2.5 Short Channel Effects

2.5.1 End Effects and Channel Length Modulation

Under the *charge sharing model*, the depletion charge induced in the body of an MOSFET is assumed to be balanced by equal but opposite charges in the gate and the depletion regions within the source and drain. The depletion region may therefore be split into three sections: one which is balanced by the gate charge and two which are balanced by the source and drain depletion charge (as shown in figure 2.11). For a long channel device, the effect of the source and drain on the depletion charge directly underneath the gate is negligible and may be ignored.

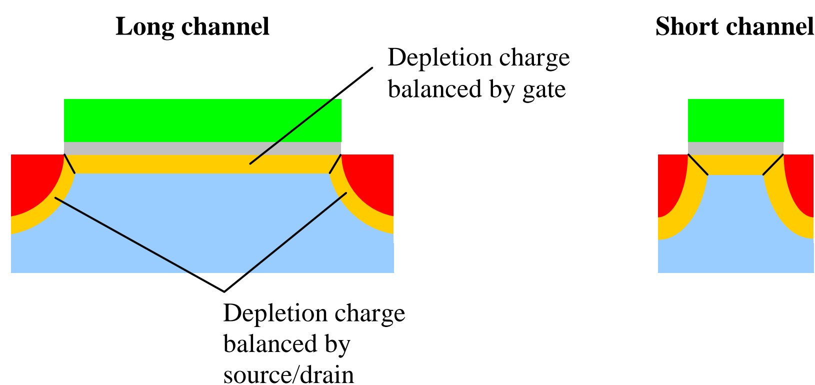


Figure 2.11. Diagram showing the splitting of depletion charge into three regions under the charge sharing model.

As the length of the channel is reduced, the depletion regions within the source and drain balance proportionally more of the charge underneath the gate. Less gate charge is now required to bring about the onset of inversion, and the threshold voltage is reduced [32] by as much as 200mV. This *threshold voltage roll-off* at short channel lengths may be combated by an increase in the body doping concentration in the vicinity of the source and drain implants.

Channel length modulation (CLM) is a reduction in the effective channel length caused by the movement of the pinch-off point away from the drain as V_{ds} is increased beyond the saturation point. Whilst this also occurs for long channel devices, the effect is only noticeable as the channel length is reduced, since the length of the pinched off region becomes significant compared to the separation of the source and drain.

2.5.2 Drain Induced Barrier Lowering

Drain induced barrier lowering (DIBL) is perhaps best described as a reduction in gate control over the channel current. In short devices the drain bias depletion region has the effect of the lowering the potential barrier at the source, with a reduction in threshold voltage and a corresponding increase in current, even when the device is in its off-state. For long channel devices the potential barrier is flat across most of the device and a dip at the drain end has negligible impact on the barrier height at the source. DIBL is another short channel effect that may be alleviated by increasing the doping concentration in the channel region.

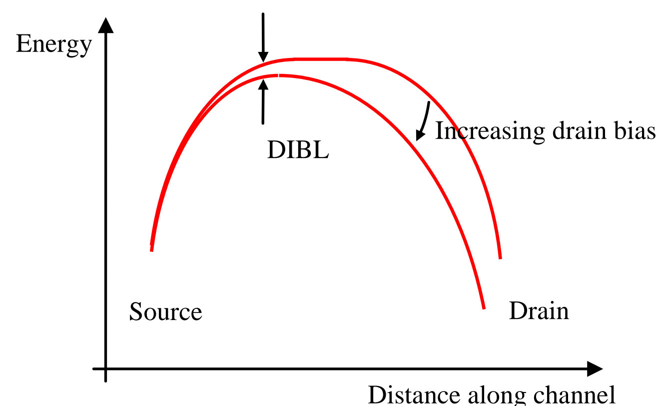


Figure 2.12. The reduction in barrier height at the source end of the channel as a bias is applied to the drain.

2.5.3 Punchthrough

Punchthrough is an extreme version of DIBL and occurs when the depletion regions of the source and drain begin to overlap. It may take the form of either surface punchthrough or bulk punchthrough, depending on the doping profile in the device. The effect is the same as having a depletion region all the way along the channel created by the gate. When a bias is applied between the source and drain, substantial current flows with no bias applied to the gate.

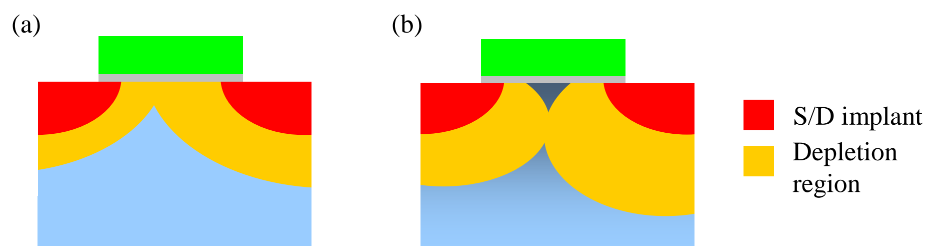


Figure 2.13. (a) Surface punchthrough and (b) bulk punchthrough in the presence of a retrograde doping profile.

As already mentioned, the solution to all the SCEs discussed so far is usually to increase the level of doping in the channel region. For a pMOSFET, if the n-type doping concentration in the bulk is increased then the size of the depletion regions around the source and drain are reduced and undesirable SCEs are diminished. Care must be taken with increasing doping in the region of the channel however, as increased scattering at ionised impurities and an elevated vertical electric field can degrade mobility and hence device performance. In order to circumvent this problem, a process known as halo doping may be used.[(33), (34)] Dopants are implanted at an angle of approximately 7 degrees from the vertical using the gate stack as a mask. The wafer is rotated whilst implantation is in progress so that a ring or "halo" of increased n-type doping is introduced just below the channel, in close proximity to the

source and drain. This has the effect of reducing the size of the depletion regions and hence SCEs, without greatly impeding carrier transport in the channel.

2.5.4 Velocity Saturation

As the length of the channel is decreased, the longitudinal electric field (parallel to the direction of current flow) increases for a given V_{ds}. The average carrier energy also increases but there are more interactions with the lattice, resulting in a reduction of carrier mobility. *Velocity saturation* starts to occur when carrier energy exceeds the minimum longitudinal optical phonon energy. Carriers begin to lose energy by optical phonon emission, which for holes occurs at a field of ~ 7×10^3Vcm^{-1} (at 300K).(35) When the longitudinal field reaches ~ 10^5Vcm^{-1}, holes travel at their saturation velocity (v_{sat}) of approximately 7×10^6cms^{-1} (36) as they lose kinetic energy as fast as they gain it from the field. The saturation for electrons occurs at lower fields because of their inherently higher mobility, although their saturation velocity is slightly higher. The addition of strain to silicon is not expected to alter the saturation velocity significantly.(37)

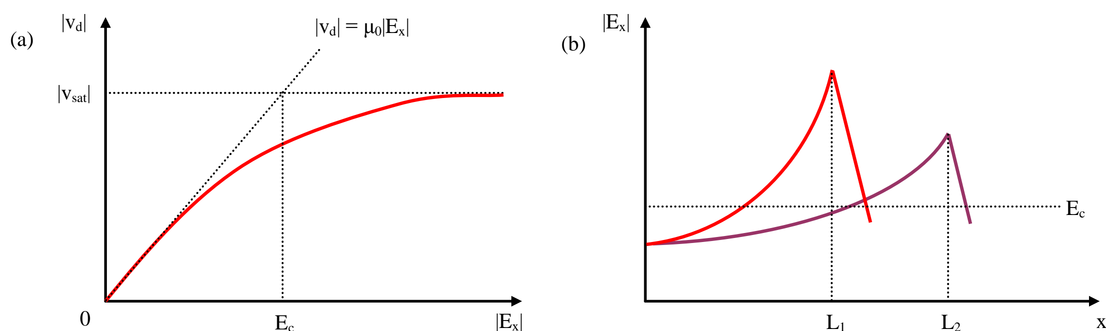


Figure 2.14. (a) Variation in carrier drift velocity with electric field and (b) variation of electric field strength along the channel of two devices with different channel lengths (after Tsividis).(38)

In short devices with non-uniform lateral doping profiles (or longer devices with a high longitudinal field) the gradual channel approximation can no longer be considered valid and carrier mobility varies along the channel. Consequently, when mobility is extracted from the electrical characteristics of short devices, it is an average value along the length of the channel, known as the *effective mobility*.

2.5.5 Velocity Overshoot

In very short devices the high lateral fields present mean that the assumption of drift-diffusion models, which is that carrier transport is always in thermal equilibrium with the lattice, breaks down. This is because there is actually a small but finite relaxation time attributed to carriers travelling through a crystal lattice, such that in a high field it is possible for carrier velocity to temporarily exceed the saturation velocity.(39) This phenomenon is called *velocity overshoot*. The peak carrier velocity attained by these *hot carriers* is approximately given by the relation $v = \mu_0 E$, so if the low lateral field mobility, μ_0, can be improved then so can velocity overshoot characteristics (figure 2.15).(40)

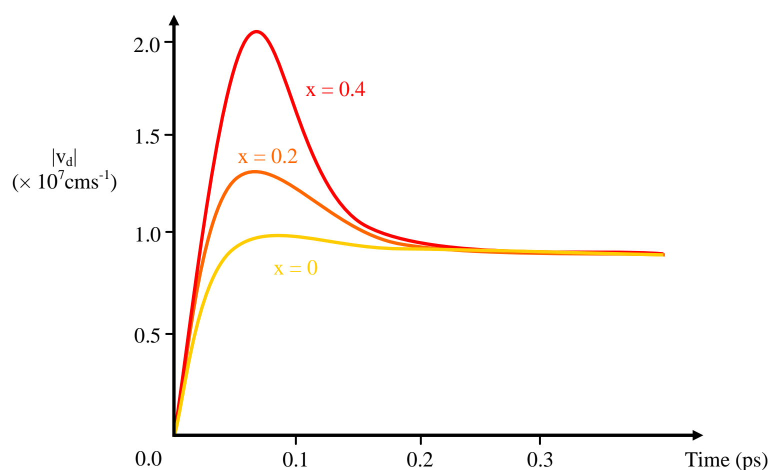


Figure 2.15. Simulated hole drift velocity as a function of time in strained silicon on $Si_{1-x}Ge_x$, and unstrained silicon (after Bufler and Fichtner).(40)

Because the highest fields in an MOSFET are found in the vicinity of the drain, if velocity overshoot is to occur then it will be in this region. At first sight, it appears as though velocity overshoot may lead to greatly enhanced device characteristics and there is interest in maximising the drain current of an MOSFET by exploiting it. However, Lundstrom [41] argues that velocity overshoot is of secondary importance to the backscattering coefficient at the source end of the channel. He shows that under saturated conditions, the drain current is given by:

$$I_{ds} = C_{ox} W v_T \left(\frac{1 - r_c}{1 + r_c} \right) \left(V_{gs} - V_t \right), \tag{2.10}$$

where the effective carrier velocity, v_{eff}, is given by:

$$v_{eff} = v_T \left(\frac{1 - r_c}{1 + r_c} \right). \tag{2.11}$$

Here v_T is the Richardson thermal velocity and r_c is the reflection coefficient, governing the proportion of carriers injected from the source into the channel that are backscattered into the source. The thermal velocity is the average speed at which carriers move through a crystal (in random directions) in the absence of an electric field. It may be 1.5 - 2 times greater than v_{sat} in a heavily doped degenerate source.[42] Monte Carlo simulations show that once an electron has travelled approximately $2kT/q$ down the potential drop to the drain (the region in which the first 50mV of the potential drop across the channel occurs) then even if it backscatters, it is extremely unlikely to return to the source. Whilst there may be considerable velocity overshoot in the channel, r_c and therefore the drain current are controlled by carriers that have only been heated by a small amount.

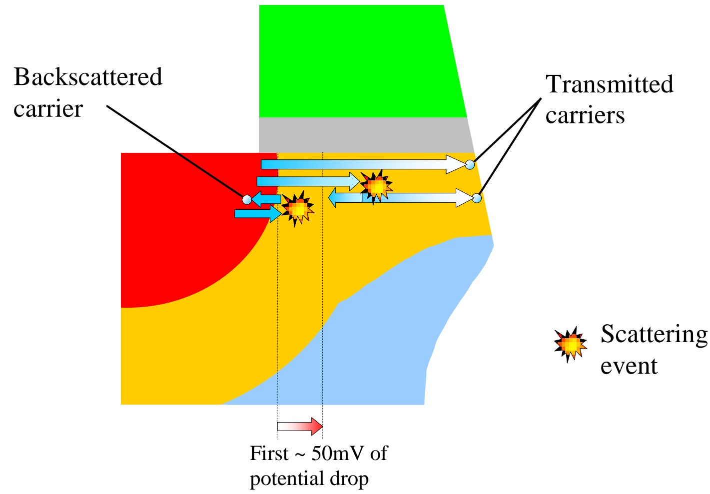


Figure 2.16. Diagram of the source region of an MOSFET. Carriers that travel beyond the initial potential drop are unlikely to return to the source.

Despite the fact that drain current is ultimately governed by the *thermal limit* because carriers can not be injected from the source to the channel any faster than the Richardson thermal velocity, velocity overshoot still has an effect on the self-consistent electric field in the channel. This in turn influences the field at the source and affects the backscattering coefficient. If modern devices were already close to the thermal limit then only a minor drive current benefit could be expected from further scaling or by introducing a higher mobility material such as strained silicon. Lochtefeld *et al.* [43] show by simulation that the commonly used expression for carrier velocity:

$$v_{eff} = \frac{g_{mi}}{WC_{ox}}, \tag{2.12}$$

where g_{mi} is the peak intrinsic transconductance, actually gives the velocity of carriers a considerable distance from the source, explaining why some authors have reported $v_{eff} > v_T$ in their devices. They propose the relation:

$$v_{idi} = \frac{I_{ds}}{WQ_i(x_0)}, \tag{2.13}$$

as a means of extracting the carrier velocity within a few nanometres of the source (v_{idi} is the intrinsic velocity extracted from the drain current). Here, $Q_i(x_0)$ is the inversion charge sheet density at the source end of the channel. Using this technique, they demonstrate that the effective carrier velocity close to the source of their $L_{eff} \approx$ 40nm MOSFETs is less than half of the thermal velocity. Additionally, it is shown that a low field mobility enhancement continues to be of importance in increasing v_{eff} in deep sub-100nm nMOSFETs.

2.5.6 Further High Field Effects

The hot carriers created by the large electric fields in short devices can cause degradation of the FET action. Some carriers may acquire enough energy for *impact ionisation* of lattice atoms to occur, liberating extra electrons and holes in a weak avalanche. Minority carriers created in this way join the stream flowing to the drain, whereas majority carriers are forced into the substrate by the normal depletion field, giving rise to a substrate current. This current induces a potential difference across the channel-substrate junction, which can charge the semiconductor bulk and reduce the threshold voltage, giving an additional increase in the drain current. This is the so-called substrate current induced body effect (SCBE), and is particularly significant when the substrate is highly resistive. If the longitudinal field exceeds the breakdown field (~ 3×10^5Vcm^{-1} for silicon, depending on doping) then avalanche breakdown occurs and a plasma of energetic electron-hole pairs is created in the surface of the depletion layer. Some of these carriers have sufficient energy to overcome the barrier of the Si/SiO_2 interface and are injected into the oxide.(44) These are then collected by the gate as gate current. Some will create damage at the Si/SiO_2 interface, seen as an increase in interface states density, and some will be trapped in the oxide, which leads

to a gradual corruption of the oxide with time known as *ageing*. Avalanche breakdown is not a necessary prerequisite for carrier injection; a small minority of hot carriers may gain enough energy for injection even in the presence of smaller fields. Aside from reliability concerns, impact ionisation is an important consideration for high frequency power amplification functions because it determines the maximum power that a technology can deliver.[45]

A possible solution to hot carrier effects is to employ a lightly doped drain (LDD) structure, in which a drain extension is created at a doping level one or two orders of magnitude lower than the drain itself. This allows more of the depletion region to exist within the drain and reduces the maximum electric field in the channel. However, in a standard fabrication process the source must also have a lightly doped extension due to symmetry and a high series resistance for the device can result.

2.5.7 Quantum Effects

Whilst hot carrier generated gate current is not usually a problem, as device dimensions are scaled down direct quantum mechanical tunnelling through the oxide becomes a serious concern.[46] Power supply voltages are also scaled down, so the drain potential is reduced and hot carrier effects are somewhat diminished. Gate potential is similarly reduced, so in order to maintain the number of carriers in the channel from one generation to the next, the capacitance of the gate stack must be increased. The simplest way of doing this is to reduce the thickness of the oxide layer, but modern devices already employ oxides of approximately 1.5nm thickness. For such thin oxides the potential barrier between channel and gate is very narrow and quantum mechanical tunnelling from the channel to the gate can result. An alternative

to reducing the oxide thickness is to replace the silicon dioxide with another material having a higher dielectric constant, but finding a suitable material that forms a high quality interface with silicon is proving to be a very difficult task. For very short devices (less than ~ 10nm in length), it is expected that direct tunnelling between the source and drain will start to become a problem.(47)

Even if quantum mechanical tunnelling were not a problem, there is a lower limit on the separation between the charge in the gate and inversion charge in the channel due to the phenomenon of quantum confinement. At the semiconductor surface, inversion charge is confined to an approximately triangular potential well, which will be very steep when the vertical field is large, as in a state-of-the-art device. The energy levels in a narrow well are spaced further apart than in a wide well, with the result that very few inversion charge carriers may occupy the region directly adjacent to the gate oxide. Instead, the resultant wave function has a peak 2 - 3nm from the semiconductor/oxide interface.

2.6 Growth of Virtual Substrates

The creation of a strained silicon layer typically involves the use of a virtual substrate, which is a region of relaxed SiGe alloy grown on top of a standard silicon wafer. In industry this is usually accomplished using chemical vapour deposition (CVD) although molecular beam epitaxy (MBE) may also be used, particularly in semiconductor research.

2.6.1 CVD

There are various CVD techniques available for the growth of SiGe heterostructures, all employing gaseous species as the source of silicon and germanium adatoms. Usually, silane (SiH_4) is used as the source of silicon and germane (GeH_4) as the source of germanium. The wafer is heated so that these 'precursors' decompose on impact, releasing mobile silicon and germanium atoms.

A common form of CVD is ultra high vacuum CVD (UHVCVD),[(48)] in which the background pressure (before the addition of precursors) is reduced to ~ 10^{-10}mbar, greatly reducing the incorporation of contaminants. The entire growth chamber is heated in order to achieve a high degree of thermal uniformity and, because the growth rates are highly temperature dependent, a high uniformity of deposition. This, however, has the disadvantages of a slow temperature response and deposition on the sidewalls of the chamber. The relative growth rates of silicon and germanium are controlled by varying the partial pressures of the source gases. Doping is typically accomplished using B_2H_6 (p-type) and PH_3 or AsH_3 (n-type), with n-type dopants suffering severely from excessive surface segregation and therefore profile abruptness and incorporation problems. The main advantage of CVD over other growth systems is the high throughput of wafers possible due to fast growth rates and multi-wafer loading.

Low energy plasma enhanced CVD (LEPECVD) [(49)] employs an inert plasma (typically argon) to crack the reactants near to the surface of the wafer. This eliminates the requirement for the wafer to be very hot since the silicon and germanium adatoms are produced in the plasma. Growth rates as high as 5nms^{-1} are

attainable using this system. Much of the material for the work carried out in this study was grown by LEPECVD.

2.6.2 MBE

MBE differs from CVD in that the epitaxial material arrives at the wafer in the form of an atomic or molecular beam. A VG Semicon V90S system was used to grow some of the material for this work (figure 2.17). This system uses an electron beam to melt a solid charge of silicon or germanium, hence this method of epitaxy is known as solid source MBE (SS-MBE).[50] The flux of silicon and germanium is controlled by the intensity of the electron beam, which requires frequent recalibration as the charges deplete. The low flux rates of SS-MBE (~ $1Ås^{-1}$) compared to those typically encountered in CVD enables the heterostructures to be controlled to a single atomic layer. As with UHVCVD, the growth chamber must be kept under UHV conditions to keep any contaminants incorporating onto the substrate to a minimum, and to ensure that the flux of silicon and germanium can reach the substrate without being involved in collisions. Uniform coverage of the wafer is achieved by rotating it at approximately 5rpm. A heater mounted above it can control the substrate temperature between room temperature and approximately 1000°C. An important feature of SS-MBE to note is that, in contrast to most forms of CVD, temperature and growth rates can be controlled independently.

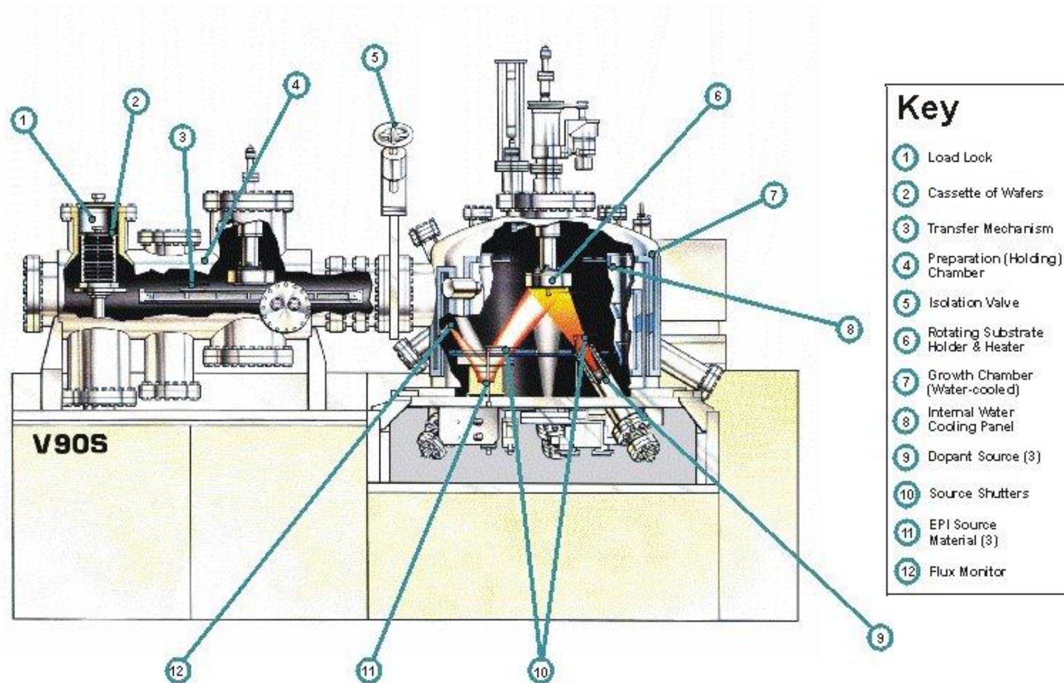


Figure 2.17. Schematic of the VG Semicon V90S MBE growth system used for the growth of some of the layers in this work.

2.6.3 Dislocations

The creation of a biaxial tensile strained silicon layer requires a region of SiGe, relaxed so that its lattice constant is sufficiently greater than that of bulk silicon. The simplest way to achieve this is by the growth of a SiGe layer of constant composition, and in excess of the critical thickness, onto a silicon substrate. As the strain energy increases with the thickness of the layer, it becomes energetically favourable for the strain to be relieved by the formation of interfacial misfit dislocations. This is usually by the nucleation of a half loop from the surface, which extends by glide to the interface between the SiGe and silicon substrate.(51) Nucleation will either be homogenous or heterogeneous. Homogeneous nucleation is the spontaneous formation of a dislocation within the semiconductor material.(52) This mechanism has high activation energy and consequently occurs when the strain mismatch is high.

Heterogeneous nucleation is the formation of a dislocation from a source that is extrinsic to the semiconductor crystal. This may be a surface contaminant due to poor cleaning, a metallic or carbide precipitate within the layer, or some other foreign material. The threading segments formed by nucleation then glide away from each other leaving a misfit dislocation between them (figure 2.18). Ideally these would travel to the edge of the wafer, leaving no threading components. However, in reality it is often the case that in order to fully relax a SiGe layer involves the formation of approximately 10^5 - 10^6 dislocations per square centimetre.(53) Since many dislocations are needed to relax the layer, they tend to interact with each other and may become pinned. This leaves threading dislocations penetrating the surface and more dislocations must be formed in order to continue the relaxation. The threading dislocation density will increase to as much as 10^9 - 10^{12}cm^{-2} for constant composition SiGe virtual substrates,(54) the upper limit being reached when pure germanium is grown on silicon.

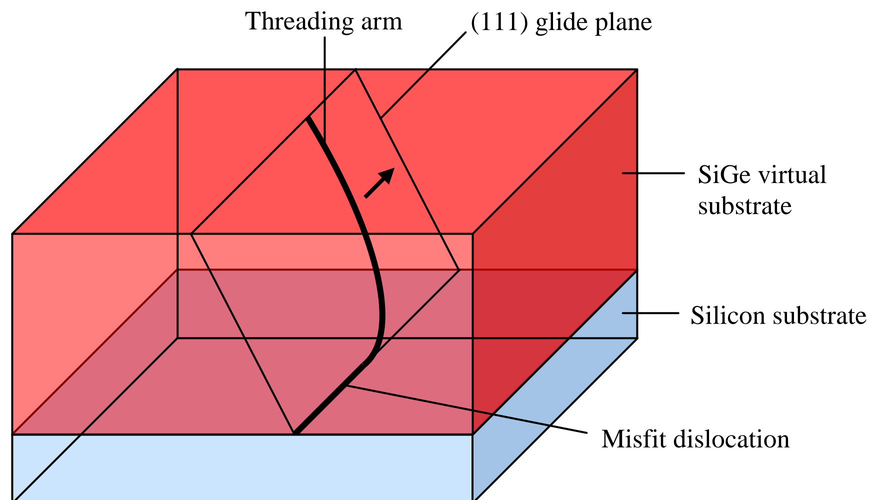


Figure 2.18. After nucleation a threading dislocation glides, leaving a misfit dislocation at the SiGe/Si interface. This dislocation relieves the strain of the region above it.

It is of paramount importance to create virtual substrates with as low a threading dislocation density as possible, in order to suppress a number of undesirable effects (see section 2.6.4). Since each misfit dislocation has two threading segments (one at each end), it is advantageous for misfits to be long. In order to accomplish this, the glide velocity of the threading sections must be increased to reduce the chance of pinning. This may be achieved by increasing growth temperature and employing a low grading rate from pure silicon to the final desired composition of SiGe.

The most common way of reducing pinning events is to linearly grade the germanium composition during growth. The idea behind this is to ensure that the dislocations are distributed over the entire width of the graded region instead of being confined to the single atomic plane of the SiGe/Si interface, thus reducing the chance of interactions. Growth temperature is usually reduced with increasing germanium content in an attempt to suppress surface roughening, due the higher adatom mobility of germanium.[55] Linear grading techniques are frequently effective at reducing the threading dislocation density of the virtual substrate to approximately 10^6cm^{-2}.[56]

2.6.4 Difficulties Associated with MOSFETs on Virtual Substrates

Devices fabricated in layers grown on virtual substrates with a high threading dislocation density can have very high substrate leakage currents, largely due to p - n leakage from the drain/body junction. Junction leakage at the drain end of the channel is higher for strained silicon on SiGe devices than standard silicon devices even in the absence of threading dislocations, due to the smaller bandgap of SiGe. This makes it easier for electrons to tunnel through the potential barrier by a process known as the Zener effect,[57] resulting in an elevated leakage current (figure 2.19).

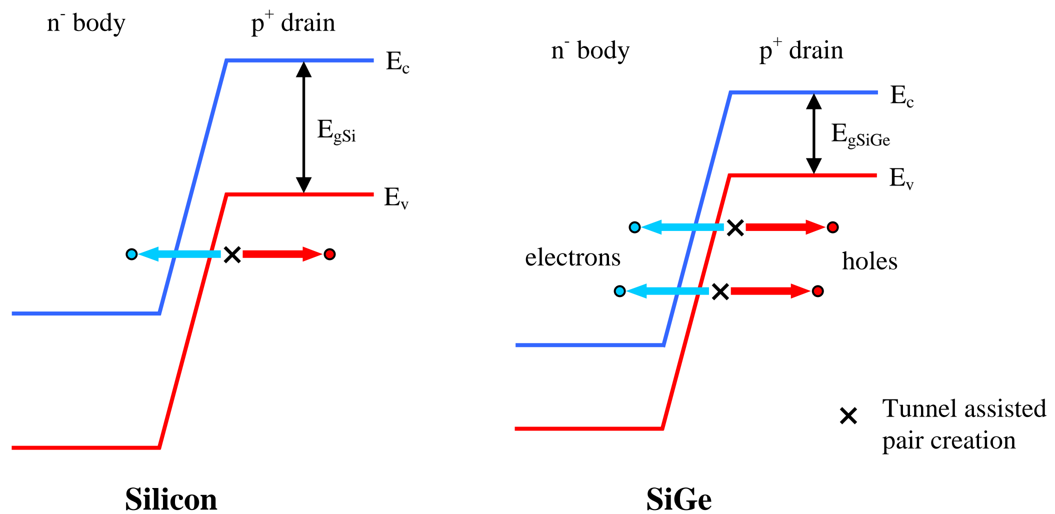


Figure 2.19. The Zener effect for reverse-biased p-n junctions. The smaller bandgap of SiGe results in elevated leakage current.

Threading dislocations in the drain region provide a path across the potential barrier via traps, resulting in a greatly increased leakage current. Devices created on virtual substrates with high threading dislocation densities will therefore tend to have poor I_{on}/I_{off} ratios, which would be unacceptable in a CMOS circuit.

A further source of leakage current experienced when an MOSFET is in its off-state is *gate induced drain leakage* (GIDL).(58) The carriers responsible for GIDL originate in the region of the drain that is overlapped by the gate. For a small device, a large electric field exists across the gate oxide in this region, such that an inversion layer will attempt to form at the semiconductor surface. In a pMOSFET, the electrons formed in the drain will then tunnel to the lower potential substrate, creating an additional leakage current and increasing standby power. Again, threading dislocations assist this process, highlighting the need for low defect virtual substrates.

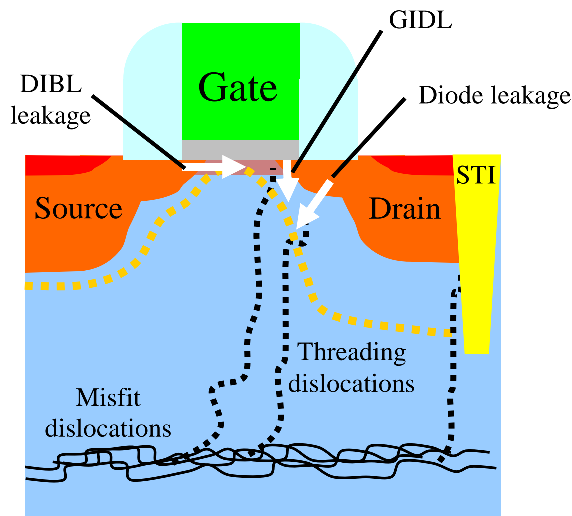


Figure 2.20. Diagram showing the leakage currents experienced by a strained silicon MOSFET in the off-state.

A pile-up of misfit dislocations associated with the modified Frank-Read mechanism (59) can result in a network of substantial surface steps known as *crosshatch*. A virtual substrate suffering from crosshatch can lead to poor carrier mobilities in layers grown on top of it. At first sight, it is not immediately apparent why this should be so, since the roughness related to crosshatch has a wavelength of approximately 1μm. This is much larger than the wavelengths of carriers in the channel and therefore should not interfere with carrier transport. Carrier mobility is thought to be degraded by the increased micro-roughness encountered at non-zero vicinal angles (60) (the angle between the surface normal and the silicon [001]) as shown in figure 2.21. A wafer with a large degree of crosshatch will have larger vicinal angles and therefore increased micro-roughness, leading to a reduction in carrier mobility.

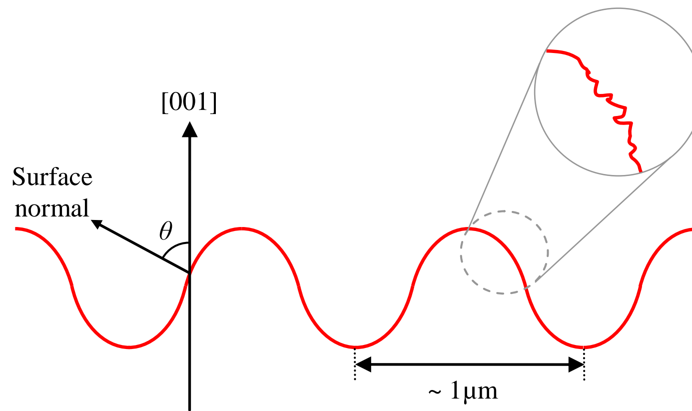


Figure 2.21. Diagram showing the relationship between vicinal angle, θ, and micro-roughness. The amplitude of the crosshatch roughness has been much exaggerated.

Crosshatch is usually reduced by keeping the grading rate low. In this way the surface is kept as far from the underlying misfit dislocation networks as possible. This reduces the effect of the strain fields associated with these networks to a minimum at the surface of the wafer. A low grading rate also implies a thick epitaxial layer, which gives a smoothing effect on the surface. There are problems associated with a thick virtual substrate, however. Firstly, a thick region of SiGe can greatly inhibit heat flow away from the channel, resulting in the phenomenon of self-heating as discussed in section 2.7. Secondly, the growth time required to create a layer of several microns thickness can be prohibitive, particularly for MBE which suffers from low growth rates and source depletion. Finally, the step height introduced by a thick virtual substrate can lead to difficulties with integrating the wafer into existing silicon CMOS fabrication facilities. A compromise grading rate is therefore required. Fitzgerald *et al.* (53) have shown that the optimal grading rate is approximately 10% germanium per micron.

2.6.5 Chemical Mechanical Polishing

A technique for reducing the surface roughness of virtual substrates, which is gaining popularity, is chemical mechanical polishing (CMP). First introduced by Currie *et al.*,[61] the virtual substrates are linearly graded at a rate of 10%μm^{-1}, as described in section 2.6.3, but the growth is interrupted at the halfway point and the wafer removed from the growth chamber for a CMP step. This removes the crosshatch before the wafer is reloaded into the chamber for the continuation of growth. The existing threading dislocations are free to glide again and there is no need for new dislocations to be nucleated. The threading dislocation density of the virtual substrate is thus reduced. Currie *et al.* find that the line density is reduced from more than 10^7cm^{-2} down to 2×10^6cm^{-2} for a virtual substrate with a final composition of 100% germanium. The rms roughness is approximately halved to 24nm. However, cross-hatch may return at elevated temperature, as the surface responds to the strain field associated with the underlying dislocation network.

2.7 Self-Heating

2.7.1 Origin and Effects of Self-Heating

Employing a virtual substrate, as described in section 2.6, to create a strained silicon cap layer inevitably means that there is a relatively thick region of SiGe directly underneath the active region of an MOSFET fabricated on such a wafer. The thermal conductivities of silicon, germanium, SiGe and SiO_2 are shown in table 2.1.[62], [63]

Material	Thermal conductivity at 300K ($Wm^{-1}K^{-1}$)
Si	150
$Si_{0.75}Ge_{0.25}$	8.5
$Si_{0.5}Ge_{0.5}$	8.3
$Si_{0.25}Ge_{0.75}$	11
Ge	60
SiO_2	1.4
Al	230
Poly Si (doped with $10^{19}cm^{-3}$ B)	45.3

Table 2.1. Thermal conductivities of the most common materials employed in a Si/SiGe MOSFET.

The thermal conductivity of SiGe varies somewhat with composition, but is in general at least fifteen times lower than that of silicon (figure 2.22). This is due to the random alloy of silicon and germanium atoms scattering phonons, which are responsible for heat transfer. The virtual substrate therefore effectively thermally insulates the device from the substrate, with the result that the power dissipated by the device in its on-state can cause self-heating.(64) The result of this is a lower than expected drain current, due to the reduced mobility of strained silicon at higher temperature.

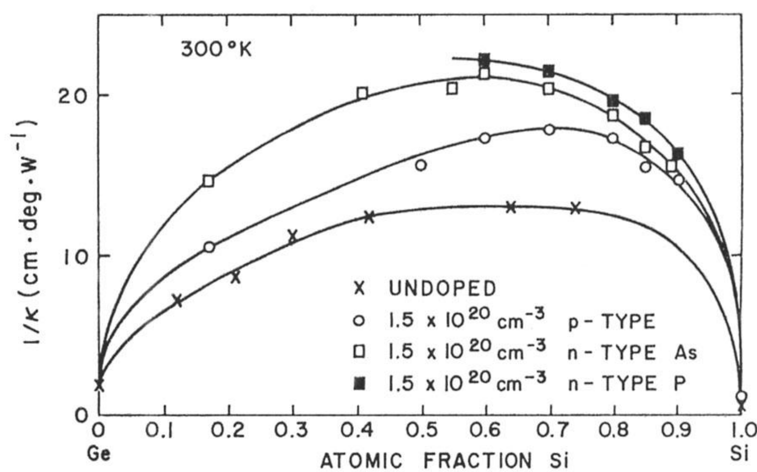


Figure 2.22. The thermal resistivity of undoped and doped, p- and n-type SiGe alloys as a function of alloy composition (after Dismukes *et al.*).(65)

As yet there has been no consideration of the effect of threading dislocations in the virtual substrate on self-heating. Kotchetkov *et al.* [66] find that the effect of threading dislocations on gallium nitride is to reduce the thermal conductivity substantially, but only when the line density approaches $10^{11}cm^{-2}$. This is because imperfections and impurities in the lattice scatter acoustic phonons, introducing additional thermal resistance. The thermal conductivities of gallium arsenide and gallium nitride are approximately 55 and $130Wm^{-1}K^{-1}$ respectively. The similarity of the two systems means that this mechanism is likely to be repeated in SiGe and so it is possible that the thermal conductivity of virtual substrates with high threading dislocations densities will be even lower than expected. It should be noted, however, that a good quality virtual substrate (with a final germanium content of 50% or more) will have a threading dislocation density five or six orders of magnitude lower than that which causes a decrease in the thermal conductivity of gallium nitride.

The issue of self-heating has previously been encountered in silicon-on-insulator (SOI) MOSFETs, which are fabricated in a thin silicon film on top of a layer of SiO_2. This yields complete dielectric isolation and allows the shrinking of distances between single devices. Additionally the source and drain junction capacitance is almost entirely eliminated, since the capacitance through the thick buried oxide layer to the substrate is very small.[67] However, owing to the poor thermal conductivity of the oxide layer, self-heating becomes an issue.

In digital circuits, MOSFETs are required to switch very rapidly. Under these conditions self-heating of SOI or strained silicon devices is not thought to be a major problem [68] because power dissipation typically only occurs during switching and

there is negligible steady state power consumption. The channel temperature does not follow the device power dissipation instantaneously, due to the finite thermal capacity of the device.[(69)] However, in analogue applications where devices are constantly sinking DC power, an integrated circuit employing SOI or strained silicon MOSFETs will have a higher operating temperature than its silicon equivalent due to self-heating, unless design steps are taken to ensure that there are sufficient thermal conduction paths. There are only three ways for thermal energy to leave the MOSFET: through the substrate, through the material on top of the device (typically protective gas) or through the interconnects. Berger and Chai [(70)] show that heat loss to the protective gas (usually air) is minimal, even taking convection into consideration. They also show that about 14% of heat transport takes place through the aluminium contacts of a typical SOI device (L = 1.6μm), with the rest being lost through the substrate. For this reason thermal vias may be used in devices that are prone to self-heating to increase heat flow through the substrate. Special consideration may also be given to the design of the source/drain and gate contacts, together with the interconnects, to increase the proportion of heat lost through these regions.

Analogue applications or characterisation of a device suffering from self-heating under DC conditions may easily result in substantial reduction of the drain current when compared to the dynamic characteristics. There are a number of methods that may be employed to obtain device characteristics in the absence of self-heating, which are discussed in the following sections.

2.7.2 Pulsed Measurements

When the problem of self-heating was originally realised, the room temperature characteristics of a device could be approximately found by attempting to measure the temperature of the device as a function of its power and then incorporating a thermal model into a device simulation [71 - 73]. This method obviously has limitations since it relies on approximations and specific device architectures. More recently, Jenkins and Sun [68] introduced the concept of pulsed I_{ds}-V_{ds} measurements, whereby the drain is left continually biased but short 7ns pulses with a low duty cycle are applied to the gate. Because the thermal time constants associated with self-heating are longer than this, the drain current is measured using a digital oscilloscope before self-heating takes effect. By performing pulsed measurements on a heated device, Jenkins and Sun determine that self-heating is responsible for a 100K temperature rise in their SOI device. They also find that, provided the duty cycle is kept low (less than 0.1%), the drain current is not reduced until the gate pulse duration is increased to several hundred nanoseconds.

The pulsed measurement technique has been widely accepted as a quick and relatively straight-forward means of extracting device parameters in the absence of self-heating. However, Tu *et al.* [74] contend that pulsed measurements may be unreliable because the source/drain bias must be adjusted to compensate for the voltage drop across the external resistor used to determine the drain current. They also argue that there may be significant heating during the initial finite rise of the pulse. However, provided the pulse is of the order of nanoseconds it has been shown that self-heating does not occur before a measurement can be taken. Additionally, provided the external resistor used

for current determination is small compared to the resistance of the device, minimal error is introduced by its inclusion and there is seldom need to adjust the applied bias.

Limitations of the pulsed technique are likely to arise due to poor quality devices. If the device under investigation has a high interface trap density then there may be a difference between pulsed and DC measurements even for a device which should not exhibit self-heating. This is because there is insufficient time during an applied bias pulse for generation-recombination events at surface states to occur and the device is measured as though the interface traps were not present, in a similar fashion to performing a high frequency C-V measurement (section 3.5.10). Perhaps more seriously, a pulsed measurement of a device with appreciable substrate leakage current will also display a deviation from the static measurement that is not due to self-heating. This is because a finite time is required for the establishment of leakage currents such as that associated with the drain junction. Depending on the quality of the virtual substrate, leakage currents may be quite large in strained silicon devices and care must be taken when comparing pulsed measurements with those obtained under DC conditions that any variation in drain current is actually due to elimination of self-heating effects and not leakage currents or interface states.

2.7.3 AC Conductance Technique

The method favoured by Tu *et al.* is the AC conductance technique. They employ an LCR meter to measure the output conductance and, by integrating with respect to drain bias, obtain the familiar I_{ds}-V_{ds} curve. The MOSFET is held under constant bias conditions but a high frequency small amplitude AC signal is applied to the drain by the LCR meter. If the frequency of the signal is high enough then the MOSFET

temperature does not respond to the signal and it is assumed that the temperature is constant. The DC self-heating effect is still present and the output resistance of the device depends on temperature. From consideration of the following equation:

$$\frac{dI_{ds}}{dV_{ds}} = \frac{\partial I_{ds}}{\partial T} \cdot \frac{\partial T}{\partial V_{ds}} + \left. \frac{\partial I_{ds}}{\partial V_{ds}} \right|_{T_0}, \tag{2.14}$$

Tu *et al.* argue that the first term on the right hand side of this equation is much larger than the change in the second term due to an increase in device temperature. The first term is removed by measuring at high frequencies and hence most of the effect of self-heating is eliminated.

However, Tenbroek *et al.* (69) argue that this is an over-simplification and is inaccurate even for small temperature rises. They offer a rather more comprehensive treatment which removes the appeal of the AC conductance technique as a simple means of device characterisation, so this is not pursued any further in this work.

References

1. Y. Taur and T. H. Ning, “Fundamentals of Modern VLSI Devices”, Cambridge University Press, pp. 112-139, 1998.
2. D. A. Neamen, “Semiconductor Physics and Devices: Basic Principles”, Richard D. Irwin Inc., pp. 530-534, 1992.
3. Y. Tsividis, “Operation and Modeling of the MOS Transistor”, 2nd Edition, WCB/McGraw-Hill, pp. 248-301, 1999.
4. Y. Taur and T. H. Ning, “Fundamentals of Modern VLSI Devices”, Cambridge University Press, pp. 19-20, 1998.
5. M. Lundstrom, “Fundamentals of Carrier Transport”, 2nd Edition, Cambridge University Press, pp. 9-12, 2000.
6. C. Kittel, “Introduction to Solid State Physics”, 7th Edition, John Wiley & Sons, pp. 98-111, 1996.
7. M. V. Fischetti and S. E. Laux, “Monte Carlo Study of Electron Transport in Silicon Inversion Layers”, *Phys. Rev. B*, vol. 48, no.4, pp. 2244-2274, 1993.
8. S. Takagi, A. Toriumi, M. Iwase and H. Tango, “On the Universality of Inversion Layer Mobility in Si MOSFET’s: Part I - Effects of Substrate Impurity Concentration”, *IEEE Trans. Elec. Dev.*, vol. 41, no. 12, pp. 2357-2362, 1994.

9. S. M. Sze, "VLSI Technologies", 2nd Edition, McGraw-Hill Book Co., pp. 359-362, 1988.
10. T. E. Whall and E. H. C. Parker, "Si/SiGe/Si pMOS Performance - Alloy Scattering and Other Considerations", *Thin Solid Films*, vol. 368, pp. 297-305, 2000.
11. F. Schäffler, "High-Mobility Si and Ge Structures", *Semicond. Sci. Technol.*, vol. 12, pp. 1515-1549, 1997.
12. T. E. Whall and E. H. C. Parker, "SiGe Heterostructures for FET Applications", *J. Phys. D: Appl. Phys.*, vol. 31, pp. 1397-1416, 1998.
13. R. Braunstein, A. R. Moore and F. Herman, "Intrinsic Optical Absorption in Germanium-Silicon Alloys", *Physical Review*, vol. 109, no. 3, pp. 695-710, 1958.
14. D. J. Paul, "Silicon-Germanium Strained Layer Materials in Microelectronics", *Advanced Materials*, vol. 11, no. 3, pp. 191-204, 1999.
15. S. Takagi, J. L. Hoyt, J. J. Welser and J. F. Gibbons, "Comparative Study of Phonon-Limited Mobility of Two-Dimensional Electrons in Strained and Unstrained Si Metal-Oxide-Semiconductor Field-Effect Transistors", *J. Appl. Phys.*, vol. 80, no. 3, pp. 1567-1577, 1996.
16. Y. H. Xie, "SiGe Field Effect Transistors", *Materials Science and Engineering*, vol. 25, pp. 89-121, 1999.
17. C. Kittel, "Introduction to Solid State Physics", 7th Edition, John Wiley & Sons, pp. 197-216, 1996.
18. T. Tezuka, A. Kurobe, N. Sugiyama and S. Takagi, "Experimental Evidence of Valence Band Deformation due to Strain in Inverted Hole Channel of Strained-Si pMOSFETs", *Thin Solid Films*, vol. 369, pp. 338-341, 2000.
19. M. T. Currie *et al.*, "Carrier Mobilities and Process Stability of Strained Si n- and p-MOSFETs on SiGe Virtual Substrates", *J. Vac. Sci. Technol. B*, vol. 19, no. 6, pp. 2268-2279, 2001.
20. K. Rim *et al.*, "Strained Si CMOS (SS CMOS) technology: opportunities and challenges", *Solid-State Electronics*, vol. 47, pp. 1133-1139, 2003.
21. M. Yang *et al.*, "High Performance CMOS Fabricated on Hybrid Substrate with Different Crystal Orientations", *IEDM Tech. Dig.*, 2003.
22. R. Oberhüber, G. Zandler and P. Vogl, "Subband Structure and Mobility of Two-Dimensional Holes in Strained Si/SiGe MOSFET's", *Phys. Rev. B*, vol. 58, no. 15, pp. 9941-9947, 1998.
23. M. V. Fischetti, F. Gámiz and W. Hänsch, "On the Enhanced Electron Mobility in Strained-Silicon Inversion Layers", *J. Appl. Phys.*, vol. 92, no. 12, pp. 7320-7324, 2002.
24. G. Abstreiter, H. Brugger, T. Wolf, H. Jorke and H. -J. Herzog, "Strain Induced Two-Dimensional Electron Gas in Selectively Doped Si/Si_xGe_{1-x} Superlattices", *Phys. Rev. Lett.*, vol. 54, no. 22, pp. 2441-2444, 1985.
25. C. K. Maiti, L. K. Bera and S. Chattopadhyay, "Strained-Si Heterostructure Field Effect Transistors", *Semicond. Sci. Technol.*, vol. 13, pp. 1225-1246, 1995.
26. M. M. Rieger and P. Vogl, "Electronic-Band Parameters in Strained $Si_{1-x}Ge_x$ Alloys on $Si_{1-y}Ge_y$ Substrates", *Phys. Rev. B*, vol. 48, no. 19, pp. 14276-14287, 1993.
27. S. B. Samavedam *et al.*, "Relaxation of Strained Si Layers Grown on SiGe Buffers", *J. Vac. Sci. Technol. B*, vol. 17, no. 4, pp. 1424-1428, 1999.

28. J. W. Matthews and A. E. Blakeslee, “Defects in Epitaxial Layers: I. Misfit Dislocations”, *J. Cryst. Growth*, vol. 27, pp. 118-125, 1974.
29. J. C. Bean, L. C. Feldman, A. T. Fiory, S. Nakahara and I. K. Robinson, “Ge_xSi_{1-x} /Si Strained Layer Superlattice Grown by Molecular Beam Epitaxy”, *J. Vac. Sci. Technol. A*, vol. 2, pp. 436-440, 1984.
30. K. Rim, J. Welser, J. L. Hoyt and J. F. Gibbons, “Enhanced Hole Mobilities in Surface-Channel Strained-Si p-MOSFETs”, *IEEE IEDM Tech. Dig.*, pp. 517-520, 1995.
31. T. J. Grasby, *private communication*.
32. M. Conti and C. Turchetti, “On the Short-Channel Theory for MOS Transistor”, *IEEE Trans. Elec. Dev.*, vol. 38, no. 12, pp. 2657-2661, 1991.
33. H. Wakabayashi *et al.*, “Sub-50-nm Physical Gate Length CMOS Technology and beyond using Steep Halo”, *IEEE Trans. Elec. Dev.*, vol. 49, no. 1, pp. 89-95, 2002.
34. W. -K. Yeh and J. -W. Chou, “Optimum Halo Structure for Sub-0.1µm CMOSFETs”, *IEEE Trans. Elec. Dev.*, vol. 48, no. 10, pp. 2357-2362, 2001.
35. Y. Taur and T. H. Ning, “Fundamentals of Modern VLSI Devices”, Cambridge University Press, pp. 149-153, 1998.
36. Y. Taur *et al.*, “Saturation Transconductance of Deep-Submicron-Channel MOSFETs”, *Solid-State Electron.*, vol. 36, no. 8, pp. 1085-1087, 1993.
37. J. B. Roldan, F. Gamiz, J. A. Lopez-Villanueva and J. E. Carceller, “Understanding the Improved Performance of Strained $Si/Si_{1-x}Ge_x$ Channel MOSFETs”, *Semicond. Sci. Technol.*, vol. 12, no. 12, pp. 1603-1608, 1997.
38. Y. Tsividis, “Operation and Modeling of the MOS Transistor”, 2nd Edition, WCB/McGraw-Hill, pp. 280-287, 1999.
39. G.G. Shahidi, D. A. Antoniadis and H. I. Smith, “Electron Velocity Overshoot at Room and Liquid Nitrogen Temperatures in Silicon Inversion Layers”, *IEEE Elec. Dev. Lett.*, vol. 9, no. 2, pp. 94-96, 1988.
40. F. M. Bufler and W. Fichtner, “Scaling and Strain Dependence of Nanoscale Strained-Si p-MOSFET Performance”, *IEEE Trans. Elec. Dev.*, vol. 50, no. 12, pp. 2461-2466, 2003.
41. M. Lundstrom, “Scattering Theory of the Short Channel MOSFET”, *IEDM Tech. Dig.*, pp. 387-390, 1996.
42. Y. Taur and T. H. Ning, “Fundamentals of Modern VLSI Devices”, Cambridge University Press, pp. 153-154, 1998.
43. A. Lochtefeld, I. J. Djormehri, G.Samudra and D. A. Antoniadis, “New Insights into Carrier Transport in n-MOSFETs”, *IBM J. Res. & Dev.*, Vol. 46, no. 2/3, pp. 347-357, 2002.
44. E. H. Nicollian and J. R. Brews, “MOS (Metal Oxide Semiconductor) Physics and Technology”, Bell Telephone Laboratories Inc., pp. 495, 1982.
45. N. S. Waldron *et al.*, “Impact Ionisation in Strained-Si/SiGe Heterostructures”, *IEEE IEDM Tech. Dig.*, 2003.
46. P. -F. Wang, T. Nirschl, D. Schmitt-Landsiedel and W. Hansch, “Design of High Performance Esaki-Tunneling FET”, *3rd European Workshop on the Ultimate Integration of Silicon*, vol. 3, pp. 103-106, 2002.
47. A. Asenov, A. R. Brown and J. R. Watling, “Quantum Corrections in the Simulation of Decanano MOSFETs”, *3rd European Workshop on the Ultimate Integration of Silicon*, vol. 3, pp. 111-114, 2002.

48. B. S. Meyerson, “Low-Temperature Silicon Epitaxy by Ultrahigh Vacuum/Chemical Vapour Deposition”, *Appl. Phys. Lett.*, vol. 48, pp. 797-799, 1986.
49. C. Rosenblad, H. von Känel, M. Kummer, A. Dommann and E. Müller, “A Plasma Process for Ultrafast Deposition of SiGe Graded Buffer Layers”, *Appl. Phys. Lett.*, vol. 76, pp. 427-429, 2000.
50. E. Kasper and J. C. Bean, “Silicon Molecular Beam Epitaxy”, CRC Press, 1988.
51. Y. Fukuda, Y. Kohama, M. Seki and Y. Ohmachi, “Generation of Misfit Dislocations in $Si_{1-x}Ge_x$ /Si Heterostructures”, *Jap. J. Appl. Phys.*, vol. 28, L19-20, 1989.
52. R. Hull, “Properties of Strained and Relaxed Silicon-Germanium”, *Emis Data Reviews*, No. 12, pp.28, 1994.
53. E. A. Fitzgerald *et al.*, “Dislocations in Relaxed SiGe/Si Heterostructures”, *Phys. Stat. Sol. (a)*, vol. 171, no. 1, pp. 227-238, 1999.
54. H. -J. Herzog, T. Hackbarth, G. Höck, M. Zeuner and U. König, “SiGe-Based FETs: Buffer Issues and Device Results”, *Thin Solid Films*, vol. 380, pp. 36-41, 2000.
55. P. M. Mooney, J. L. Jordan-Sweet, J. O. Chu and F. K. LeGoues, “Evolution of Strain Relaxation in Step-Graded SiGe/Si Structures”, *Appl. Phys. Lett.*, vol. 66, no. 26, pp. 3642-3644, 1995.
56. F. K. LeGoues, B. S. Meyerson, J. F. Morar and P. D. Kirchner, “Mechanism and Conditions for Anomalous Strain Relaxation in Graded Thin Films and Superlattices”, *J. Appl. Phys.*, vol. 71, no. 9, pp. 4230-4243, 1992.
57. D. A. Neamen, “Semiconductor Physics and Devices: Basic Principles”, Richard D. Irwin Inc., pp. 301, 1992.
58. S. Wolf, “Silicon Processing for the VLSI Era. Volume 3: The Submicron MOSFET”, Lattice Press, pp. 198-200, 1995.
59. M. A. Lutz, R. M. Feenstra, F. K. LeGoues, P. M. Mooney and J. O. Chu, “Influence of Misfit Dislocations on the Surface Morphology of $Si_{1-x}Ge_x$ Films”, *Appl. Phys. Lett.*, vol. 66, no. 6, pp. 724-726, 1995.
60. S. H. Olsen *et al.*, “Impact of Virtual Substrate Quality on Performance Enhancements in Strained Si/SiGe Heterojunction n-channel MOSFETs”, *Solid State Electron.*, vol. 47, no. 8, pp. 1289-1295, 2003.
61. M. T. Currie, S. B. Samavedam, T. A. Langdo, C. W. Leitz and E. A. Fitzgerald, “Controlling Threading Dislocation Densities in Ge on Si using Graded SiGe layers and Chemical-Mechanical Polishing”, *Appl. Phys. Lett.*, vol. 72, no. 14, pp. 1718-1720, 1998.
62. M. H. Jones and S. H. Jones, “The General Properties of Si, Ge, SiGe, SiO_2 and Si_3N_4”, *unpublished.*
63. A. D. McConnell, S. Uma and K. E. Goodson, “Thermal Conductivity of Doped Polysilicon Layers”, *unpublished.*
64. K. A. Jenkins and K. Rim, “Measurement of the Effect of Self-Heating in Strained-Silicon MOSFETs”, *IEEE Electron Device Lett.*, vol. 23, no. 6, pp. 360-362, 2002.
65. J. P. Dismukes, L. Ekstrom, E. F. Steigmeier, I. Kudman and D. S. Beers, “Thermal and Electrical Properties of Heavily Doped Ge-Si Alloys up to 1300°K”, *J. Appl. Phys.*, vol. 35, no. 10, pp. 2899-2907, 1964.

66. D. Kotchetkov, J. Zou, A. A. Balandin, D. I. Florescu and F. H. Pollak, "Effect of Dislocations on Thermal Conductivity of GaN Layers", *Appl. Phys. Lett.*, vol. 79, no. 26, pp. 4316-4318, 2001.
67. Y. Taur and T. H. Ning, "Fundamentals of Modern VLSI Devices", Cambridge University Press, pp. 280-283, 1998.
68. K. A. Jenkins and J. Y.-C. Sun, "Measurement of I-V Curves of Silicon-on-Insulator (SOI) MOSFET's Without Self-Heating", *IEEE Electron Device Lett.*, vol. 16, no. 4, pp. 145-147, 1995.
69. B. M. Tenbroek, M. S. L. Lee, W. Redman-White, R. J. T. Bunyan and M. J. Uren, "Self-Heating Effects in SOI MOSFET's and their Measurement by Small Signal Conductance Techniques", *IEEE Trans. Electron Dev.*, vol. 43, no. 12, pp. 2240-2248, 1996.
70. M. Berger and Z. Chai, "Estimation of Heat Transfer in SOI-MOSFET's", *IEEE Trans. Electron Dev.*, vol. 38, no. 4, pp. 871-875, 1991.
71. L. T. Su, J. E. Chung, D. A. Antoniadis, K. E. Goodson and M. I. Flik, "Measurement and Modelling of Self-Heating in SOI NMOSFETs", *IEEE Trans. Electron Dev.*, vol. 41, no. 1, pp. 69-75, 1994.
72. N. Rinaldi, "On the Modelling of the Transient Thermal Behaviour of Semiconductor Devices", *IEEE Trans. Electron Dev.*, vol. 48, no. 12, pp. 2796-2802, 2001.
73. A. L. Caviglia and A. A. Iliadis, "Linear Dynamic Self-Heating in SOI MOSFET's", *IEEE Electron Device Lett.*, vol. 14, no. 3, pp. 133-135, 1993.
74. R. H. Tu, C. Wann, J. C. King, P. K. Ko and C. Hu, "An AC Conductance Technique for Measuring Self-Heating in SOI MOSFET's", *IEEE Electron Device Lett.*, vol. 16, no. 2, pp. 67-69, 1995.

Chapter 3

Experimental Method and Device Analysis

3.1 Introduction

This chapter describes the experimental apparatus and methods that were employed in gathering the results presented in this work. Transmission electron microscopy (TEM), X-ray diffractometry (XRD) and secondary ion mass spectrometry (SIMS) were used to glean structural information about the heterostructures under investigation whilst current-voltage (I-V) and capacitance-voltage (C-V) measurements were used in electrical characterisation of devices. Pulsed I-V measurements were also made, using a system developed by the author, and are described in chapter 4.

3.2 TEM

3.2.1 Sample Preparation

In order that enough electrons can pass through a sample to form an image in the TEM, it is necessary to make the sample very thin (100nm or less). This is accomplished primarily by mechanically grinding small pieces of a wafer down to a thickness of approximately 20μm, with further thinning achieved by ion beam milling. The procedure is very laborious and, owing to the accuracy required, frequently unsuccessful.

Figure 3.1 illustrates the technique that was used for sample preparation. In brief, two pieces of approximately 4 × 10mm were cleaved from a wafer using a diamond scribe. The uppermost sides were glued together and silicon support blocks glued to each underside. After curing in an oven, the sample was cut perpendicular to its

longest side to create three smaller samples, using a diamond saw. The purpose of this was to reduce the necessary grinding and to increase the chances of successfully producing a finished foil. The samples were attached to a metal block using beeswax and ground using progressively finer grits and polished on a Metaserv 2000 wheel. The samples were then turned and the process repeated, with the objective of reducing the thickness to the required 20µm. A copper ring of 3mm diameter was glued onto each sample and, after removal from the block and cleaning in hot propan-2-ol, the foils were milled in a Gaton PIPS ion beamer. Argon ions were accelerated using a potential of 4.5kV at an angle of 4° to the plane of the foil. It was evident that the sample was suitable for examination under the TEM when a small hole began to appear in the centre, with the material surrounding the hole being electron transparent.

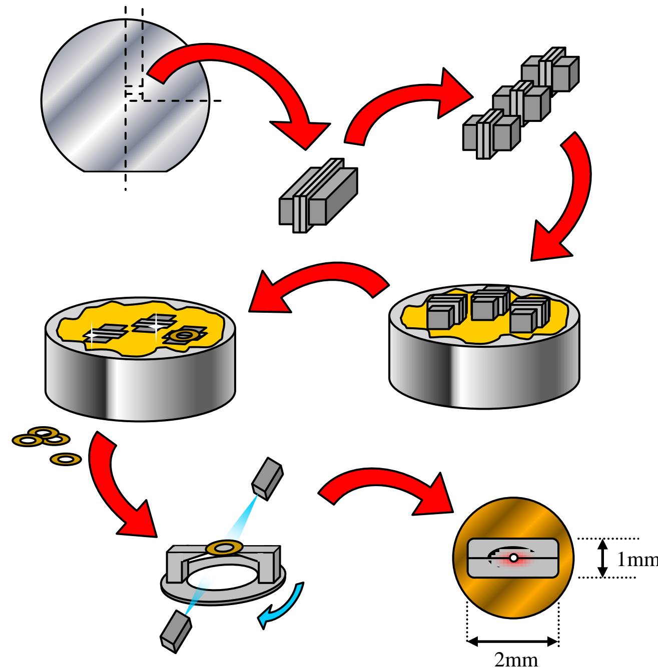


Figure 3.1. The procedure for preparing samples to be viewed with the TEM.

3.2.2 Image Formation

The cross-sectional TEM (X-TEM) images presented in this work were obtained using a JEOL JEM-2000FX TEM. A schematic diagram of the microscope is presented in figure 3.2. A uniform beam of electrons is created by an electron gun, consisting of a tungsten wire filament connected to the cathode of a high voltage supply, a grounded anode and a Wehnelt cap. The filament is heated to produce thermionic emission of electrons, which are then accelerated towards the anode by the high voltage supply (set at the maximum of 200kV for this work). The Wehnelt cap shields the gun and has a focussing effect on the beam. Further control of the electron beam is gained by use of the condenser aperture, which has the effect of filtering out the most wayward electrons. The lenses in the TEM are high quality electromagnets, designed to produce a homogeneous magnetic field. Having passed through the condenser aperture, the electron beam encounters the two condenser lenses, the purpose of which is to control the spot size and the intensity of the electrons at the sample.

Because electrons interact strongly with matter, the entire TEM column is kept under high vacuum conditions, typically $< 10^{-7}$mbar, by a combination of molecular diffusion and ion pumps. Isolation valves between the different chambers within the system allow for the loading and unloading samples without venting the column. The sample to be investigated is held in a specimen holder which permits the sample to be translated and rotated in any direction. This allows any part of the sample to be viewed in the desired crystallographic orientation.

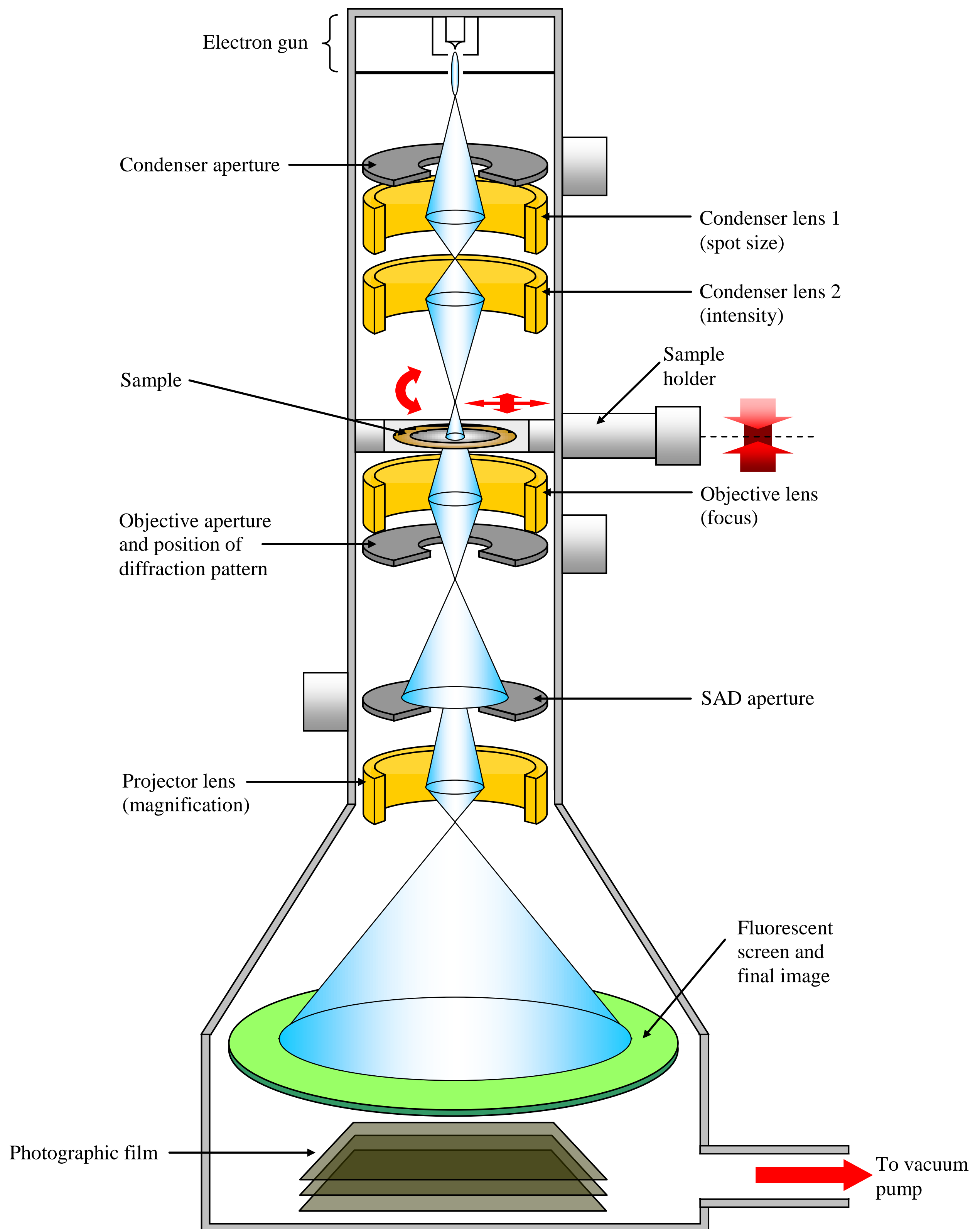


Figure 3.2. Schematic diagram showing the main components of the TEM.

Immediately after passing through the sample, the electron beam encounters the objective lens and aperture. The objective lens is the primary means of controlling the focus of the final image and also produces a real image of the sample at approximately fifty times magnification. The objective aperture is used to select specific diffraction conditions, which are important in producing contrast in the TEM (see section 3.2.3). The selective area diffraction (SAD) aperture is positioned where the diffraction pattern image is formed. In imaging mode this aperture is removed from the beam, but in diffraction mode it is used to select the area of the sample allowed to contribute to the diffraction pattern. The final lens is the projector lens, which provides the majority of the magnification and produces a final image on a fluorescent phosphor screen. The screen may be moved to allow photographs to be taken.

3.2.3 Contrast in TEM Images

Contrast in the TEM is supplied mainly by the objective aperture. This is accomplished by preventing the continuation of the majority of diffracted beams, produced by the interaction of the original electron beam with the sample, to the formation of the final image. The TEM images in this work are *bright-field* images, meaning that it is the un-diffracted beam that forms the image. *Dark-field* images are sometimes favoured to improve contrast when features are indistinct. Two diffraction conditions are usually favoured when investigating crystalline samples with the diamond structure (as is the case with silicon, germanium and alloys of the two), which have been grown in the [001] direction. In X-TEM these are the <004> and <220> directions; the former being used for investigation of layer thickness and interface quality and the latter for the viewing of dislocations.

Contrast is produced when there is strong diffraction from a set of planes with varying spacing because the intensity of the diffracted beam changes with the spacing. In a pseudomorphically strained layer the largest change occurs in the [001] growth direction. The microscope is therefore set up such that only the (000) and (004) spots are strongly excited and the objective aperture is positioned to exclude the contribution from the diffracted beam in the formation of a bright-field image. It should be noted that the (004) spot is excited because the (001), (002) and (003) spots are all forbidden in a diffraction pattern of silicon.

Contrast due to dislocations is greatest in a direction parallel to the Burgers vector [1] and as such when viewing dislocations, the TEM should be set up in such a way that the (220) spot is excited. This is because the expected Burgers vector of a dislocation will be in the [110] direction for silicon/germanium and the (110) diffraction spot is forbidden.

3.3 SIMS

SIMS measures isotopic composition as a function of depth using a beam of primary ions (in this case O_2^+). The sample under investigation is eroded as the beam is scanned across a square (of width ~ 200µm) of the surface and neutral atoms and secondary ions are ejected. The secondary ions account for only approximately 1% of the particles coming from the surface and are identified in a mass spectrometer, measuring their mass to charge ratio. The concentration of a given ion as a function of time is therefore recorded and subsequently converted to concentration versus depth data by measuring the depth of the erosion crater after the measurement and assuming a constant erosion rate. Digital gating may be employed in such a way that only the secondary ions produced from the centre of the crater are recorded. This

effectively removes the problem of secondary ions being ejected from the sides of crater, which can greatly reduce resolution.

The resolution of SIMS is limited by the fact that the primary ions are normally able to penetrate the sample to a depth of a few nanometres, thus at any given moment the mass spectrometer will be recording the concentration of ions over this thickness of the sample. The Atomika - 4500 SIMS Profilometer used in this work utilises low beam energies (as low as 75eV is possible, compared to the more usual 1 - 25keV [(2)]), which improves the resolution at the expense of erosion rate. Low energy reduces channelling and enables the ion beam to strike the sample at normal incidence. Uneven surfaces will also create a broadening of the SIMS profile on the scale of the roughness amplitude. A further limitation is the ability to measure the secondary ions sputtered from the sample. The mass spectrometer used is able to detect concentrations down to approximately 5×10^{16}cm^{-3}, although this is dependent on species.

3.4 XRD

By diffracting X-rays from a semiconductor sample, it is possible to determine the composition of that sample and the degree of strain that is present. The condition for the diffraction of a wave is given by the Bragg equation:

$$2d_{hkl} \sin\theta = \lambda , \qquad (3.1)$$

where d_{hkl} represents the spacing of the planes with Miller indices h, k and l, and θ is the angle of diffraction measured from the hkl plane. A strained layer grown on a substrate will be tetragonally distorted, such that the spacing of the layer will not generally be the same as that of the substrate. By measuring the in plane and out of

plane lattice constants (a_x and a_z), the relaxed lattice constant may be found, if it is assumed that the material is elastic and the Poisson's ratio, v, is known:

$$a = a_x + \left(\frac{1-\nu}{1+\nu}\right)(a_z - a_x)\,. \qquad (3.2)$$

Since there may be a tilt of the sample with respect to the holding stage and also a tilt of the crystallographic planes of the layer being examined, measurements are always taken with reference to the Bragg angle of the silicon substrate, which is well defined. The tilt angles are determined by rotating the sample between measurements.

Once the lattice constant is known, the composition of layer can be calculated by numerically solving the corrected form of Vegard's law:

$$a(x) = a_{Si}(1-x) + a_{Ge}x + 0.027x^2 - 0.027x\,, \qquad (3.3)$$

where x is the fraction of germanium present. The state of relaxation follows from a comparison of the relaxed lattice constant and the measured in plane lattice constant.

3.5 Electrical Characterisation

3.5.1 Equipment and Measurement Techniques

In order to determine the important performance parameters of a given MOSFET, a number of different I-V and C-V measurements were made and subsequently analysed. I-V measurements were made using one of two parameter analysers: an Agilent 4156C or a Hewlett-Packard 4145B. Some quasi-static C-V measurements were also made using the Agilent 4156C, but high frequency C-V measurements required the use of a Keithley 590 CV Analyzer and 230 Programmable Voltage Source combined via a 5951 Remote Input Coupler. This system was capable of

performing C-V measurements at both 100kHz and 1MHz in addition to quasi-static measurements.

The device under test (DUT) was typically on a wafer, which was mounted in an earthed Faraday cage before characterisation. In this way electromagnetic interference was eliminated, including visible light which can generate electron-hole pairs in the active regions of a device. Devices were characterised using a four point probe method; the four contacts being the insulated gate, the source, drain and the substrate. Needle probes with a fine positioning control were used (together with a microscope) to contact directly to the pads of the device. In the event that the processing stages had included creation of a substrate contact on the top of the wafer then this was sometimes used, but more commonly the substrate connection was made to the metal plate on which the wafer was placed. A small membrane vacuum pump ensured that the wafer was unable to move relative to the plate. Connection between the probes and the test equipment was by means of triaxial and coaxial cables.

Care had to be taken in order to avoid destroying thin gate oxides with electrostatic discharges. An MOSFET with a thin oxide (~ 3nm) could easily be destroyed by a discharge of just a few volts, although some MOSFETs had much thicker oxides and could withstand ± 50V, which was the maximum the voltage source could offer. Devices were stored in antistatic containers and an antistatic wristband (connected to the mains earth) was worn whenever wafers or chip packages were handled. When connecting to devices the substrate contact was always made first, followed by the source and drain contacts, before the gate contact was finally made. In this way it was hoped that any discharges between the equipment and the device would occur well

away from the sensitive gate oxide layer. Care was also taken to switch off the microscope lamp prior to making connections. Once measurements were complete, devices were disconnected in the reverse order to connecting them (i.e. gate contact first). It is unclear how many of these precautions were strictly necessary but since it required minimal effort to adhere to them, it would have been foolhardy not to do so.

3.5.1.1 I-V Measurement Technique

There are two main types of I-V measurement. The first and perhaps most useful configuration is to ground the source and substrate and vary the gate bias, V_{gs}, whilst maintaining a constant source to drain bias V_{ds}. The drain current, I_{ds}, is then measured as a function of V_{gs}, from which a number of characteristics may be deduced. The transconductance plot follows from a simple differentiation of I_{ds} with respect to V_{gs}, and the threshold voltage may be calculated using the method outlined in section 3.5.2. Combined with quasi-static C-V measurements, this type of I-V measurement is also used to extract carrier mobility as a function of gate voltage. This information is usually determined from measurements taken at low V_{ds} (typically ± 50mV), since under these conditions the conducting channel has a relatively constant electric field along its length and consequently a constant carrier population. Measurements with larger V_{ds} are useful in determining the expected device performance were it to be employed in a CMOS circuit, where both the drain and gate are biased at the level of the power supply voltage.

By plotting drain current logarithmically, this measurement configuration also gives the subthreshold swing (section 3.5.3). The limits of the drive current (I_{on}) and the off-state leakage current (I_{off}) are also revealed, allowing a simple calculation of the

I_{on}/I_{off} ratio. By repeating the measurement for different V_{ds} values, it is possible to extract the DIBL (section 3.5.4).

The second measurement configuration holds the gate at a constant potential with respect to the grounded source and substrate and measures the drain current in response to a variation in V_{ds}. When one such measurement is complete, V_{gs} is changed and the process repeated to build up a family of curves. Such curves are useful from the point of view of revealing if the device in question is suffering from short channel effects or self-heating, since for any applied V_{gs} it is expected that the drain current will saturate once V_{ds} exceeds $V_{gs} - V_t$, as given by equation 2.5. This configuration also allows the drive current for a given supply voltage to be compared between different MOSFETs, as this is a crucial indicator of performance. It is for this reason that such measurements are carried out subsequent to a determination of threshold voltage, V_t, (section 3.5.2) since a fair comparison between devices requires that the same *gate overdrive*, $V_{gs} - V_t$, is used in each case.

3.5.1.2 C-V Measurement Technique

The outputs of the CV analyzer and voltage source were combined by the input coupler and connected to the device gate contact. This equipment was controlled by a personal computer running a program written using Capital Equipment Corporation's Testpoint^{TM} control software. The program was originally written by Dr Martin Prest of the nano-silicon group at Warwick but has been modified by most members of the group over the years. The standard C-V measurement is a two contact technique (gate and substrate). During a quasi-static measurement, the capacitance meter applies a series of bias steps to one terminal of the device and integrates the transient current

response detected on the other terminal. During a high frequency measurement the meter measures the response to an oscillating potential bias. The voltage source was only necessary if the ± 20V range supplied by the CV analyzer proved insufficient.

3.5.2 Threshold Voltage

The threshold voltage, V_t, is loosely defined as the value of V_{gs} at which the MOSFET begins to turn on. This is an important parameter whenever devices are to be compared and features in many of the equations that describe MOSFET operation. More rigorously, the threshold voltage is that value of V_{gs} required for the difference between the Fermi energy and the intrinsic Fermi energy at the surface to be equal in magnitude but opposite in sign to the difference between them in the semiconductor bulk (figure 3.3).[3] This is very difficult to measure experimentally and consequently there are many alternative definitions of the threshold voltage which lend themselves to measurement more easily.

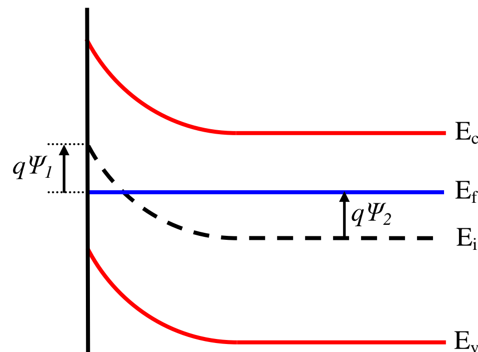


Figure 3.3. Band bending in n-type silicon under a negative applied gate bias. The device is at threshold when $\Psi_1 = \Psi_2$.

The earliest and simplest definition of threshold voltage was to choose a value of drain current (typically 100nAμm^{-1}) and say that the MOSFET was on when the

current exceeded this value.[4] Whilst this method has the attraction of simplicity, for the majority of cases it is in fact an oversimplification since it assumes that device operation scales with geometry and that properties such as the carrier mobility, which affect the drain current, are the same in all devices. Additionally it does not account for leakage currents when the device is in its off-state.

A more common technique is the *linear extrapolation method*.[3] The I_{ds}-V_{gs} curve of the device is obtained as described in section 3.5.1.1 with low V_{ds} to ensure that the device is in the linear region. The drain current is expected be linearly dependent on applied gate voltage, as shown by equation 2.4, which is repeated here:

$$I_{ds} = \mu_{eff} \frac{W}{L} C_{ox} \left[\left(V_{gs} - V_t \right) V_{ds} - \frac{V_{ds}^2}{2} \right]. \qquad (3.4)$$

However, at low V_{gs} the drain current will exceed the predicted linear dependence on gate voltage because of subthreshold currents and, at high V_{gs}, mobility degradation effects and series resistance mean that I_{ds} will fall below this linear dependence. Between these two regimes there is a point of maximum gradient on the I_{ds}-V_{gs} curve which by this method is extrapolated back to the point of zero drain current, with this being defined as V_t. This technique is usually more reliable than the constant current method and suffers only when the tail of the I_{ds}-V_{gs} curve is long, in which case there will be appreciable current flowing through the device at applied gate voltages rather less than V_t. This is usually a problem for short channel devices.

The method used in this work is to proceed in the same manner outlined for the linear extrapolation method but instead of plotting drain current against applied gate voltage, transconductance, g_m, is plotted as a function of V_{gs} where:

$$g_m = \frac{\partial I_{ds}}{\partial V_{gs}}. \tag{3.5}$$

The point of maximum gradient in the transconductance plot is then extrapolated to zero and this is defined as V_t.[4] Because this point of maximum gradient is usually in the tail of the I_{ds}-V_{gs} curve, this method tends to obtain a threshold voltage closer to the point where I_{ds} becomes negligibly small and is the preferred technique for short channel devices. Care must be taken, however, because this method is sensitive to noise in the measurement or rounding errors caused by insufficient resolution. For consistency this method was used in the extrapolation of V_t, even for long channel devices.

This technique assumes that the mobility in the threshold region is dominated by Coulomb scattering and is therefore proportional to the inversion carrier sheet density, N_s:

$$\mu_{eff} \approx \alpha N_s \approx \frac{\alpha C_{ox}}{q}\left(V_{gs} - V_t\right). \tag{3.6}$$

Substituting into 3.4 and differentiating with respect to V_{gs} yields an expression for transconductance:

$$g_m = \frac{\partial I_{ds}}{\partial V_{gs}} = \frac{W}{L}\frac{2\alpha}{q}C_{ox}^2\left(V_{gs} - V_t\right)V_{ds}. \tag{3.7}$$

As expected, g_m exhibits a linear dependence on V_{gs} in the region of interest and extrapolation to zero gives $V_{gs} = V_t$.

Tsuno [4] also argues that this technique is valid even for large V_{ds} and a short channel. Since V_t was always determined for small V_{ds} in this work, these considerations are not presented here.

3.5.3 Subthreshold Swing

As mentioned in section 3.5.1.1, the subthreshold swing (or slope) follows directly from a plot of $\log_{10}(I_{ds})$ against V_{gs}:

$$S = \left(\frac{d(\log_{10} I_{ds})}{dV_{gs}} \right)^{-1} = 2.3 \frac{kT}{q} \left(1 + \frac{C_{dep}}{C_{ox}} \right), \tag{3.8}$$

which is measured in mV/decade.[5] Here C_{dep} is the areal capacitance of the depletion region, although it should be noted that equation 3.8 will be in error in the case of a high interface trap density, since the capacitance associated with interface traps is in parallel with the depletion layer capacitance (section 3.5.5), reducing the slope.

Subthreshold swing is an important characteristic of an MOSFET, because whilst the threshold voltage may be adjusted to meet the requirements of high drive current or low leakage currents, the subthreshold swing will dictate the other parameter. Importantly, it is not possible to improve upon a swing of approximately 60mV/decade at room temperature, which has severe implications for the standby power of aggressively scaled devices, operating with small power supply voltages.

3.5.4 DIBL

DIBL is determined from two $\log_{10}(I_{ds})$ vs. V_{gs} measurements with different V_{ds} and is found from:

$$\text{DIBL} = \frac{\Delta V_t}{\Delta V_{ds}}, \tag{3.9}$$

at a given drain current. The value of I_{ds} at which to make the measurement of ΔV_t is open to some debate. Wolf [6] advises measuring ΔV_t at $I_{ds} = 10^{-7}$ Aμm^{-1} for submicron

MOSFETs, and this was the technique used for this work. In truth the value of I_{ds} at which the measurement is taken does not make a great deal of difference to the value of DIBL obtained, as long as it lies within the linear region of the subthreshold plot (figure 3.4) and provided there is not degradation of the subthreshold slope with increasing V_{ds} (punchthrough).

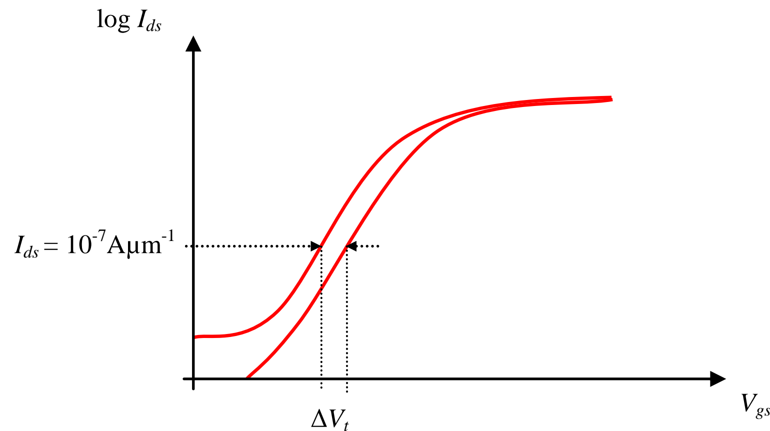


Figure 3.4. Determination of DIBL from two subthreshold curves measured at different V_{ds}.

3.5.5 Channel Length and Series Resistance

Many of the simple models for the source/drain current in an MOSFET do not take into account deviations from the designed gate length which may arise and also ignore parasitic resistances introduced at the source and drain contacts and extension implants. In order that carrier mobility may be accurately compared between devices, it is important to extract ΔL and R_{sd} from I-V measurements. Such corrections can usually be safely ignored in the case of long devices; however, as the channel length decreases to a few microns and below this is no longer valid. In this section two methods for finding ΔL and R_{sd} are outlined.[7]

The effective channel length of an MOSFET will not be the same as the expected gate length because some of the source/drain dopant may reside underneath the gate after ion implantation and diffusion during subsequent annealing steps. In addition, fringing effects at the ends of the gate electrode mean that the electric field will not be vertical along the length of the channel. Lithographic errors may also cause the written gate length to deviate from that desired. Consequently, the effective length, L_{eff}, differs from the expected length, L, by some amount ΔL, where:

$$L_{eff} = L - \Delta L \,. \tag{3.10}$$

Some of the potential V_{ds} applied across the source and drain of a device will be dropped across resistances associated with the contacts. The potential applied across the channel, V_{ds}', is given by:

$$V_{ds}^{'} = V_{ds} - I_{ds} R_{sd} \,, \tag{3.11}$$

where R_{sd} is the combined series resistance of the source and drain regions. The effect of the series resistance on MOSFET performance is particularly serious at short channel lengths. For a short device, drain current does not saturate due to CLM (section 2.5.1) and a reduction in V_{ds}' will result in a reduction in I_{ds}. Furthermore, due to the drop on the source resistance, the surface potential at the source end of the channel will be reduced and the inversion charge created by the gate bias is decreased.

The major contributor to R_{sd} is the resistance of the heavily doped semiconductor regions beneath the metal contacts, which is dependent on implant dose and activation thermal budget. In the case of an LDD structure being employed, this will add to the resistance considerably. An additional resistance results from carriers moving from this heavily doped region (which is typically some tens of nanometres thick) to a thin inversion layer as they move from the source contact to the channel. There are also

several other minor sources of contact resistance. The metal contacts themselves will have resistance, although this is negligibly small, and there is a resistance associated with the Schottky barrier which may be formed between the metal and the semiconductor surface. Again this is typically very small because the depletion region formed is very narrow in the case of a heavily doped semiconductor and carriers are relatively free to move across the barrier. Finally, there is a spreading resistance associated with current spreading out in an approximately spherical fashion from the metal contact.(8)

3.5.5.1 Linear Regression of Resistance versus Length

The measured device resistance, R_m, is given by:(9)

$$R_m = \frac{V_{ds}}{I_{ds}} = \frac{L - \Delta L}{W \mu_{eff} q N_s} + R_{sd} . \tag{3.12}$$

Typically devices used in research have many different channel lengths, allowing R_m to be plotted as a function of L for a given gate overdrive. Varying the gate overdrive produces a series of straight line graphs which, under linear regression, should all intercept at the same point where $L = \Delta L$ and $R_m = R_{sd}$.

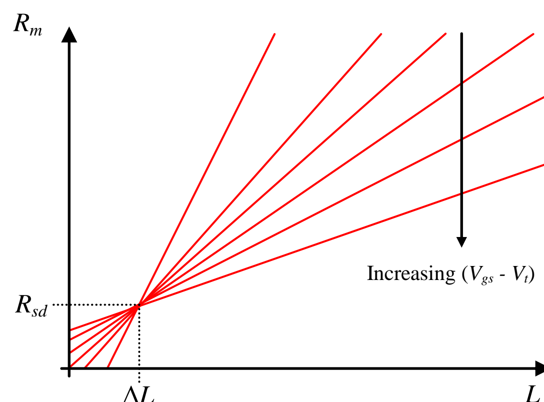


Figure 3.5. Example of the linear regression of resistance vs. length method of extrapolating R_{sd} and ΔL.

The advantage of this method is that it is a simple way of extracting the desired information. Unfortunately this method can lead to errors with a poor intercept due to R_{sd} and ΔL varying with gate overdrive, whereas it is implicitly assumed that they do not. Another drawback of the method is that because the devices must all be measured for the same gate overdrive, there is a reliance on the extrapolated threshold voltage, which may not accurately compare long and short devices. If long devices are used as part of the extrapolation process then they also have a large amount of leverage on the gradient of the line and the errors introduced in ΔL may be very large. This may be avoided by only using devices with a relatively short channel length.

3.5.5.2 Double Regression Method

An alternative method, which unfortunately suffers from many of the same problems as the simple linear regression procedure outlined in the previous section, is the double regression method proposed by Terada and Muta.[10] From equation 3.12 we have that:

$$R_m = AL + R_{sd} - A\Delta L, \qquad A = \frac{1}{W\mu_{eff} q N_S}. \tag{3.13}$$

A plot of R_m against L with gate overdrive as a parameter is created as before. The gradient A and intercept R_{sd} - $A\Delta L$ are extracted for each line by regression and then plotted against each other. This should yield a straight line with gradient -ΔL and intercept R_{sd} found by a second linear regression. Error bars for each point may be added according to how good the original linear fits were to the data.

Due to the vagaries of the methods used to extract ΔL and R_{sd}, the parameters were extracted using both techniques described here. Typically the results lay within

experimental error of each other, providing some reassurance that the values were reasonably accurate.

3.5.6 Quasi-static C-V

The purpose of a quasi-static C-V measurement is primarily to determine the areal oxide capacitance, C_{ox}, which may be used in mobility extraction. Figure 3.6 simplifies the MOSFET structure into the basic capacitances that are present.(11) In a real structure there will also be capacitances between the gate and source/drain and between the substrate and source/drain but for a device with a large area these are usually negligible. It should be noted that in order to obtain relatively "clean" C-V data, large area devices are frequently used anyway since the capacitance of smaller devices may be close to resolution limit of the equipment, resulting in noisy profiles.

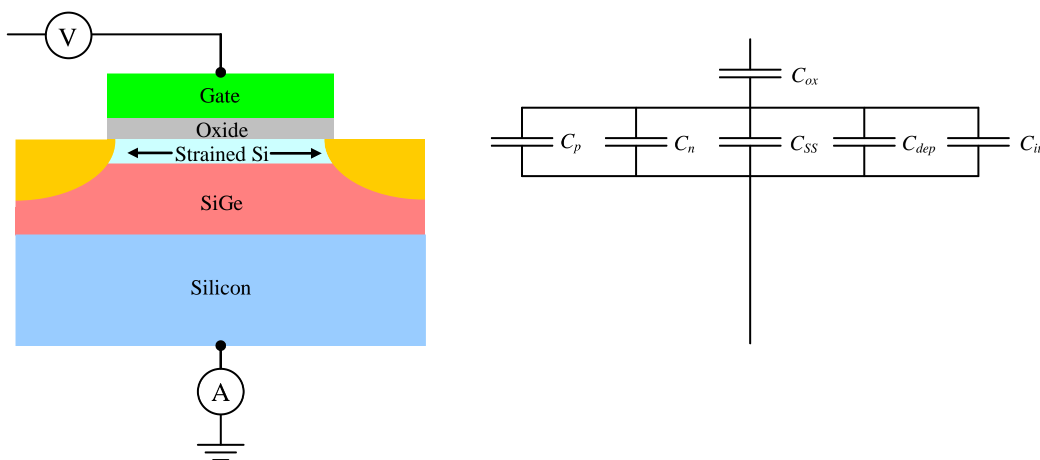


Figure 3.6. Simple structure of a strained silicon MOSFET and the major capacitances present.

It now follows that:

$$\frac{1}{C} = \frac{1}{C_{ox}} + \frac{1}{C_p + C_n + C_{dep} + C_{SS} + C_{it}}, \tag{3.14}$$

although it should be noted that not all of these capacitances are present at any one time. C_{ox} is given by:

$$C_{ox} = \frac{\varepsilon_0 \varepsilon_r}{t_{ox}}, \tag{3.15}$$

where ε_0 is the permittivity of free space, ε_r is the relative permittivity of the oxide (3.9 for SiO_2) and t_{ox} is the oxide layer thickness. C_p and C_n are the inversion and accumulation layer capacitances per unit area respectively (for a pMOSFET), C_{dep} is the areal depletion layer capacitance and C_{it} is the areal capacitance of the interface traps. C_{SS} represents the areal capacitance of the strained silicon layer in the event that one is present and hole population of the relaxed SiGe buffer has occurred.

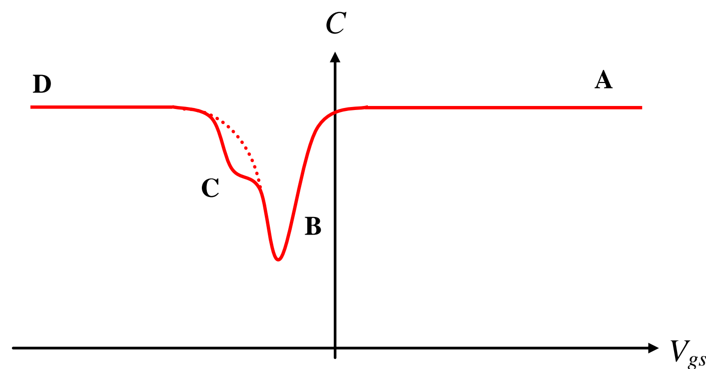


Figure 3.7. Typical quasi-static C-V characteristic of a strained silicon pMOSFET (solid line) and a standard silicon pMOSFET (dotted line).

Figure 3.7 shows the typical result of quasi-static C-V measurements on pMOSFETs. At point A there is a large positive bias applied to the gate and the device is in the accumulation region, with majority carriers (electrons) in the n-type material under the gate being strongly attracted to the oxide-semiconductor interface. The sheet density of electrons will respond to any variation in the applied gate bias and screens the bulk of the semiconductor from the gate, effectively reducing the device to a

simple parallel plate capacitor. Because the carriers are very close to the oxide, C_n is very large and dominates the second term in equation 3.14, resulting in the measured capacitance C, being very close to C_{ox}.

As V_{gs} becomes more negative, the majority carriers are pushed away from the surface and a depletion region begins to form (B). The measured capacitance is now the result of two parallel plate capacitors in series; that due to the oxide and that due to the forming depletion region. Hence:

$$\frac{1}{C} = \frac{1}{C_{ox}} + \frac{1}{C_{dep}}, \qquad (3.16)$$

$$C_{dep} = \frac{\varepsilon_0 \varepsilon_r}{t_{dep}}, \qquad (3.17)$$

where interface traps have been neglected. The relative permittivities of bulk silicon and germanium are 11.9 and 16.0 respectively. It is likely that the effect of tensile strain will be to slightly increase ε_r from the point of view of these measurements, since the thin silicon layer is compressed in the vertical direction. This may also be important when calculating vertical electric field, but as yet the variation of ε_r with strain has not been determined and the value for bulk silicon is routinely assumed. A linear variation of ε_r with germanium content is also assumed for SiGe alloys.

The depletion region increases in thickness as V_{gs} is made more negative until the point where the band bending in the semiconductor is sufficient for it to become energetically favourable for minority carriers to be created (or supplied from the source/drain regions in the case of an MOSFET) and an inversion layer forms. For a strained silicon pMOSFET in weak inversion it is possible that the inversion layer will form first in the SiGe buffer, owing to the alignment of the valence bands (C). It

is possible to estimate the thickness of the strained silicon layer if a plateau in the C-V curve forms, since the measured capacitance will be approximately given by:

$$\frac{1}{C} = \frac{1}{C_{ox}} + \frac{1}{C_{SS}} . \tag{3.18}$$

Eventually it becomes favourable for an inversion layer to form at the Si/SiO_2 interface and the interface charge becomes a linear function of applied gate bias (D).

3.5.7 Split C-V Measurements

Whereas the standard C-V method outlined in the previous section may be performed on either a capacitor or an MOSFET, the split C-V measurement introduced by Koomen [12] is a technique that may only be used on MOSFETs. The current induced in the source/drain contacts and the substrate contact may be examined separately, which allows calculation of the carrier sheet density and depletion charge sheet density.

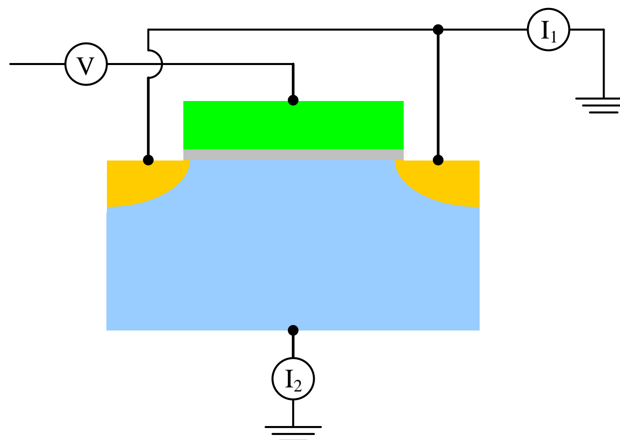


Figure 3.8. Diagram of the experimental setup for split C-V measurements.

When the current response is measured through the source and drain contacts (I_1 in figure 3.8), the measured capacitance per unit area, C_{sd}, will be approximately equal

to C_{ox} when the surface is under inversion, in the same way as for a standard quasi-static measurement. This is because the heavily doped source and drain regions are readily able to supply minority carriers to the depletion region. However, as the gate voltage sweeps towards accumulation, the measured capacitance drops to zero since it is the substrate that provides the majority carriers needed. Similarly, a capacitance measurement using ammeter I_2 will record the areal capacitance measured through the substrate, $C_{sub} \approx C_{ox}$ in accumulation, dropping to zero in inversion. A typical output of a quasi-static C-V measurement is shown in figure 3.9.

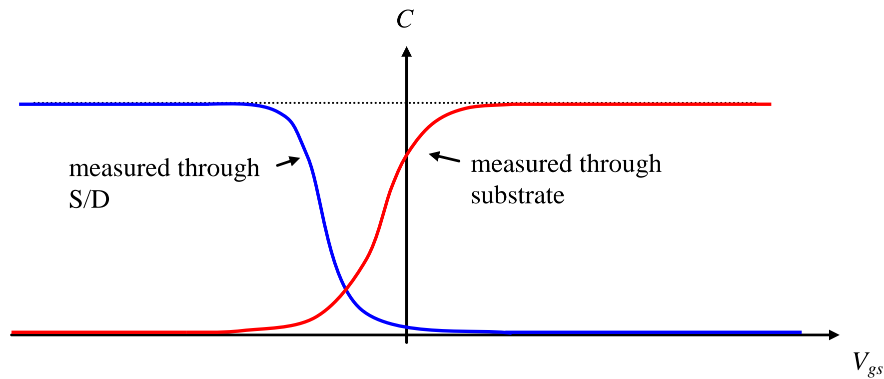


Figure 3.9. Typical result of a split C-V measurement on a silicon pMOSFET.

The inversion carrier sheet density for a given V_{gs} now follows from a simple integration of the capacitance measured through the source/drain contacts:

$$N_s(V_{gs}) = \frac{1}{q}\int_{V_{gs}}^{\infty} C_{sd}(V_{gs})dV_{gs} \ . \tag{3.19}$$

In order to find the depletion charge sheet density, it is first necessary to find the flatband voltage. This may be calculated as shown in section 3.5.10, if the doping concentration is known, or may be taken directly from a high frequency C-V measurement. A plot of $(1/C_{HF})^2$ against V_{gs} (where C_{HF} is the measured high

frequency capacitance) has a lower knee at the flatband voltage, V_{FB}.[13] The depletion carrier sheet density is obtained by integrating the split C-V curve measured through the substrate contact up to the limit of flatband:

$$N_{dep} = \frac{1}{q}\int_{-\infty}^{V_{FB}} C_{sub}(V_{gs})dV_{gs}\ , \tag{3.20}$$

for a pMOSFET.

The *effective vertical field*, E_{eff}, which is an approximation to the average normal electric field a carrier in the channel experiences, as a result of applied gate bias combined with the field supplied by depletion charge, is now given by:

$$E_{eff} = \frac{q}{\varepsilon_0 \varepsilon_r}\left(N_{dep} + \eta N_s\right). \tag{3.21}$$

Here η is an empirically determined factor, which is equal to ½ for electrons and ⅓ for holes.[14]

3.5.8 Mobility Extraction

The effective mobility may be found from the expression:

$$I_{ds} = qN_s \mu_{eff}\left(V_{ds} - I_{ds}R_{sd}\right)\frac{W}{L}\ . \tag{3.22}$$

Here the term $I_{ds}R_{sd}$ has been introduced to account for some of the applied drain bias being dropped across the source/drain resistances. The assumption made in equation 3.4 is that:

$$qN_s = C_{ox}\left(V_{gs} - V_t\right). \tag{3.23}$$

The problem with this is that it relies on an extracted V_t, which introduces some uncertainty. In addition, the prediction that the carrier sheet density is a linear function of V_{gs} is invalid in weak inversion. Nevertheless, this approximation is

usually satisfactory in the case of surface channel devices, but in the case of a strained silicon pMOSFET the relaxed SiGe buffer may also be occupied. The capacitance to the carriers is no longer C_{ox} and will vary as a function of applied gate voltage, V_{gs}. C-V allows N_s to be found directly from the measurement and obviates the need for these assumptions.

Equation 3.22 still makes the assumption that diffusion current is negligible, which is invalid close to threshold. Sodini *et al.* [15] introduce the correction:

$$\frac{I_{ds}}{W} = qN_s \mu_{eff} E_x - \frac{kT}{q} \mu_{eff} \frac{d(qN_s)}{dx}, \quad (3.24)$$

where the Einstein relation for diffusion current has been used:

$$D_p = \frac{kT}{q} \mu_{eff} . \quad (3.25)$$

From numerical simulations they also show that:

$$C(V_{gs}) = C_{ox} F(V_{gs}), \quad (3.26)$$

so that the measured capacitance approaches C_{ox} as a function of the gate voltage:

$$F(V_{gs}) = \frac{dQ_i / d\psi_s}{C_{ox} + dQ_s / d\psi_s}, \quad (3.27)$$

where Q_i and Q_s are the inversion charge and total charge in the semiconductor respectively and ψ_s is the surface potential. They also find that:

$$\frac{d(qN_s)}{dx} = \frac{C_{ox} V_{ds}}{L} F(V_{gs}). \quad (3.28)$$

Substituting this into equation 3.24 and assuming $E_x = V_{ds}' / L_{eff}$ (section 3.5.5) yields the final expression for the effective mobility that was used throughout this work:

$$\mu_{eff} = \frac{I_{ds}}{V_{ds}'} \frac{L_{eff}}{W} \frac{1}{\left(qN_s - {kT}/{q}\, C(V_{gs})\right)} . \quad (3.29)$$

3.5.9 Extraction of Doping Profile

The doping profile of the material directly underneath the gate stack determines the shape of the quasi-static C-V curve in depletion, since a higher doping concentration will result in a larger C_{dep}. It is therefore possible to approximately extract the doping profile from a C-V measurement, up to a depth determined by the onset of inversion. The method is only approximately correct because it is actually the equilibrium majority carrier density that is measured, which may differ from the ionised dopant density in the case of a non-uniform profile. This is because a non-zero field, and therefore charge neutrality, cannot exist in the presence of a non-uniform doping profile and so the carrier density does not exactly equal the dopant density.

The resolution limit of this technique is the Debye length, L_D, which is a measure of the distance over which majority carriers respond to an electric field.[16] It is a purely material related quantity and therefore constant for a given semiconductor when the doping concentration is uniform. The Debye length is given by:

$$L_D = \left(\frac{\varepsilon_0 \varepsilon_r kT}{q^2 N_D} \right)^{1/2}, \qquad (3.30)$$

where N_D is the net doping density. The resolution of this technique in the presence of a typical dopant concentration of 10^{17} - 10^{18}cm^{-3} is therefore of the order of 10nm.

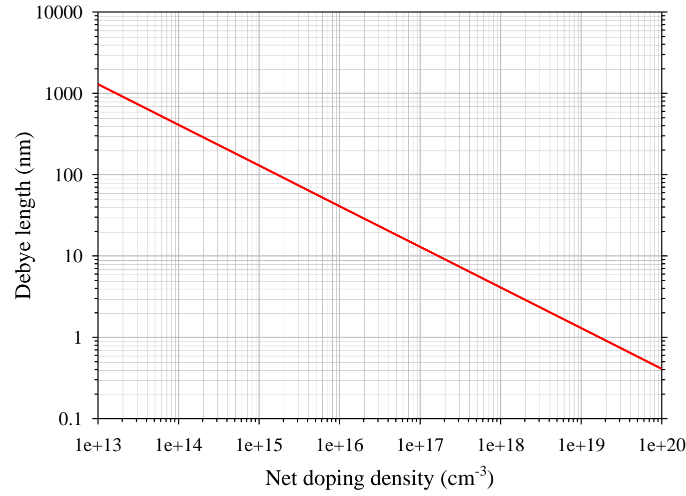


Figure 3.10. Variation of the Debye length with net doping density for silicon.

The doping concentration is related to the change in the measured capacitance, *C*, with respect to gate voltage by:

$$N(z) = \pm 2\left(q\varepsilon_0\varepsilon_r A^2 \, d\left(1/C^2\right)\Big/ dV_{gs} \right)^{-1}, \tag{3.31}$$

where the sign depends on whether the dopant distribution is p- or n-type and *A* is the device area.

3.5.10 Low and High Frequency C-V for Extraction of Interface Trap Density

This technique exploits the fact that it takes some finite time for interface traps to charge or discharge; thus if a C-V measurement is performed at a high enough frequency there will be insufficient time for many of the interface traps to respond and the measured capacitance will exclude the term C_{it} in equation 3.14.(17) Because the high frequency measurement consists of the normal slow bias ramp with a small AC signal on top (typically 1MHz), the high frequency C-V curve is still stretched out relative to that of an ideal capacitor with zero interface trap density due to the traps effectively screening the semiconductor from changes in gate charge. If capacitance

versus band bending were plotted instead of capacitance against V_{gs} then the high frequency curve would coincide with the ideal case. The existence of interface traps therefore affects the form of the high frequency C-V curve but does not contribute any additional capacitance. By comparing a high frequency C-V measurement with one performed under quasi-static conditions it is possible to extract the charge density of these interface traps as a function of position in the band gap.

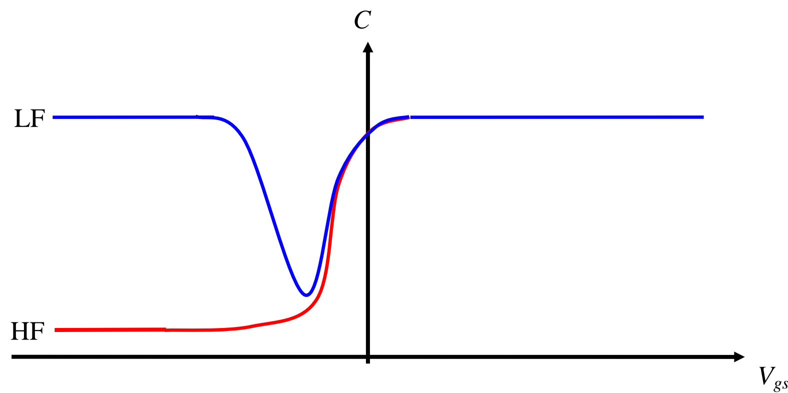


Figure 3.11. Example output characteristic of a C-V measurement performed on a pMOS capacitor at high and low frequency.

If the measurement is performed on an MOS capacitor then there are no source or drain regions to supply minority carriers for the formation of an inversion layer, and generation and recombination processes must be relied upon to source and sink the carriers. These processes have associated time constants and consequently there will be insufficient time for carrier generation to occur if the C-V measurement is carried out at high frequency, with the result that the measured capacitance does not return to C_{ox} in inversion.

In depletion we have that $C_p \approx C_n \approx 0$ and therefore:

$$\frac{1}{C_{LF}} = \frac{1}{C_{ox}} + \frac{1}{C_{dep} + C_{it}}, \tag{3.32}$$

$$\frac{1}{C_{HF}} = \frac{1}{C_{ox}} + \frac{1}{C_{dep}}, \qquad (3.33)$$

where C_{LF} and C_{HF} are the capacitances measured under quasi-static and high frequency conditions respectively. Rearranging these equations and eliminating C_{dep} yields an expression for C_{it} in terms of easily measurable quantities:

$$C_{it} = \frac{C_{ox}C_{LF}}{C_{ox} - C_{LF}} - \frac{C_{ox}C_{HF}}{C_{ox} - C_{HF}}. \qquad (3.34)$$

The high/low frequency C-V method therefore measures the interface trap sheet density, $D_{it} = C_{it}/q$, as a function of energy in the band gap. To extract this energy requires some further work. The surface potential, ψ_s, is related to the applied gate voltage in depletion by the expression:

$$\psi_s = V_{gs}\left(1 - \frac{C_{LF}}{C_{ox}}\right), \qquad (3.35)$$

where the applied gate bias has been split between that dropped between the gate and semiconductor surface and that dropped across the semiconductor itself. To relate the energy of a trap to this surface potential requires a fixed reference. This is the flatband voltage; the value of V_{gs} for which the valence and conduction bands are flat at the surface. This will be non-zero because of the influences of oxide charge and the difference in work function between the semiconductor and the gate material. The theoretical flatband capacitance is given by:

$$C_{FB} = \left(\frac{1}{C_{ox}} + \frac{L_D}{\varepsilon_0 \varepsilon_r}\right)^{-1}, \qquad (3.36)$$

where L_D is the Debye length (see section 3.5.9).

Once the flat band gate bias is known, the surface potential required for flatband, ψ_{sFB}, is calculated using equation 3.35. The band bending at a given gate voltage then

follows by subtracting the flatband surface potential from that induced by the gate bias. The energy of a trap relative to the respective band follows by adding the separation between the band edge and the intrinsic level and subtracting the separation between the Fermi level and the intrinsic level, $q\varphi_B$ (figure 3.12).

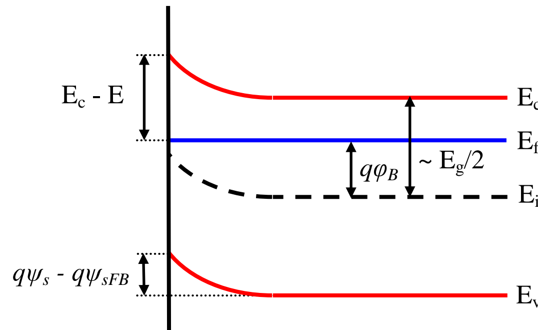


Figure 3.12. Band bending in n-type silicon under a negative applied gate bias.

The high/low frequency C-V analysis of interface trap density is only valid in the depletion region or weak inversion because of the assumption that the carrier sheet density and hence C_p is negligible.(18) In moderate or strong inversion, the increasing number of minority carriers will introduce considerable error to the technique as they are unable to respond to the high frequency measurement. There is therefore usually a steep rise in a D_{it} curve near inversion which is not due to interface traps. A similar rise is seen as accumulation is approached, as the error in the determination of C_{dep} by both the high and low frequency measurements becomes unacceptably large.

Further limitations of the high/low frequency method include the fact that at 1MHz, which is the frequency most commonly used for the C_{HF} measurement, there will still be some (greatly reduced) interface trap response. The C_{HF} curve will therefore differ from the ideal case with no interface traps by a small amount for any given band bending. Also, if the determination of the flatband capacitance requires calculation of

the Debye length, then a knowledge of the doping concentration is also required. If this is not well defined or is not constant then error will be introduced. It should be noted however, that this error will not affect the calculated values of interface trap density but only the calculated energy with respect to the band edge. In this work, the flatband voltage was usually found using the technique described in section 3.5.7. Despite its limitations, the high/low frequency method is preferable to other methods which use only a single low frequency or high frequency measurement, as these require comparison with a theoretical capacitance with no interface traps, creating considerable error.

3.5.11 Extraction of Carrier Velocity

In order to find how close to the thermal limit an MOSFET is operating, and to determine whether any velocity overshoot is occurring, it is important to know the velocity of carriers in the channel. The average carrier velocity is typically found from the peak transconductance, as mentioned in section 2.5.5. In this region of operation, the gradual channel approximation is not usually valid. The term $\mu dV_c/dx$ in equation 2.3 is replaced with an empirical expression relating carrier velocity to the horizontal electric field, E:[19]

$$v_{eff} = \frac{\mu_{eff} E}{1 + E/E_c}, \tag{3.37}$$

for holes, where E_c is the critical field ($E_c = v_{sat}/\mu_{eff}$). An alternative expression for the drain current may now be found, which in the limit as $L \rightarrow 0$ yields I_{dsat} for a short device:

$$I_{dsat} = C_{ox} W v_{eff} \left(V_{gs} - V_t \right). \tag{3.38}$$

It is now apparent that:

$$v_{eff} = \frac{g_{mi}}{WC_{ox}}, \qquad (3.39)$$

once the appropriate corrections have been made for R_{sd}.

Using equation 3.39 can result in average carrier velocities larger than the saturation velocity if significant velocity overshoot is occurring. However, it is also useful to know if carriers are being injected into the channel from the source with a velocity close to the thermal velocity, in order to gauge the potential for improvement. As discussed in section 2.5.5, a technique introduced by Lochtefeld *et al.* [(20)] allows for the extraction of carrier velocity very much closer to the source. If the inversion charge sheet density at the source can be found, then the carrier velocity at the source is given by:

$$v_{idi} = \frac{I_{ds}}{WQ_i(x_0)}. \qquad (3.40)$$

Determining $Q_i(x_0)$ in small devices is problematic due to relatively large fringing and overlap capacitances, non-uniform charge distribution along the channel, and uncertainties surrounding the effective length. However, in strong inversion and under the gradual channel approximation, $Q_i(x_0)$ in a short channel device should correspond closely to that of its long channel equivalent, which may be determined by the split C-V method (section 3.5.7). Corrections must be made to equation 3.19 to account for the difference in threshold voltage between long and short devices due to DIBL and V_t roll-off, and the potential drop across the source resistance. Accordingly, the sheet density at the source of a short channel device is given by:

$$Q_i(x_0) = \int_{V_{gs}-\Delta V_{gs}+I_{ds}R_s}^{\infty} C_{sd}(V_{gs})dV_{gs}, \qquad (3.41)$$

for a p-channel device.

References

1. A. D. Capewell, *PhD Thesis*, University of Warwick, pp. 82-84, 2002.
2. M. G. Dowsett, “Depth Profiling Using Ultra-Low-Energy Secondary Ion Mass Spectrometry”, *Applied Surface Science*, vol. 203-204, pp.5-12, 2003.
3. D. K. Schroder, “Semiconductor Material and Device Characterization”, 2nd Edition, John Wiley & Sons, pp. 242-244, 1998.
4. M. Tsuno *et al.*, “Physically-Based Threshold Voltage Determination for MOSFETs of All Gate Lengths”, *IEEE Trans. Elec. Dev.*, vol. 46, no. 7, pp. 1429-1434, 1999.
5. Y. Taur and T. H. Ning, “Fundamentals of Modern VLSI Devices”, Cambridge University Press, pp. 128-129, 1998.
6. S. Wolf, “Silicon Processing for the VLSI Era. Volume 3: The Submicron MOSFET”, Lattice Press, pp. 213-222, 1995.
7. D. K. Schroder, “Semiconductor Material and Device Characterization”, 2nd Edition, John Wiley & Sons, pp. 223-234, 1998.
8. M. J. Palmer, *PhD Thesis*, University of Warwick, pp. 78-83, 2001.
9. J. G. J. Chern, P. Chang, R. F. Motta and N. Godinha, “A New Method to Determine MOSFET Channel Length”, *IEEE Elec. Dev. Lett.*, vol. 1, no. 9, pp. 170-173, 1980.
10. K. Terada and H. Muta, “A New Method to Determine Effective Channel Length”, *Jap. J. Appl. Phys.*, vol. 18, no. 5, pp. 953-959, 1979.
11. E. H. Nicollian and J. R. Brews, “MOS (Metal Oxide Semiconductor) Physics and Technology”, Bell Telephone Laboratories Inc., pp. 71-92, 1982.
12. J. Koomen, “Investigation of the MOST Channel Conductance in Weak Inversion”, *Solid-State Electron.*, vol. 16, pp. 801-810, 1973.
13. D. K. Schroder, “Semiconductor Material and Device Characterization”, 2nd Edition, John Wiley & Sons, pp. 350, 1998.
14. S. Takagi, A. Toriumi, M. Iwase and H. Tango, “On the Universality of Inversion Layer Mobility in Si MOSFET’s: Part I - Effects of Substrate Impurity Concentration”, *IEEE Trans. Elec. Dev.*, vol. 41, no. 12, pp. 2357-2362, 1994.
15. C. G. Sodini, T. W. Ekstedt and J. L. Moll, “Charge Accumulation and Mobility in Thin Dielectric Transistors”, *Solid-State Electron.*, vol. 25, no. 9, pp. 833-841, 1982.
16. E. H. Nicollian and J. R. Brews, “MOS (Metal Oxide Semiconductor) Physics and Technology”, Bell Telephone Laboratories Inc., pp. 380-385, 1982.
17. Y. Tsividis, “Operation and Modeling of the MOS Transistor”, 2nd Edition, WCB/McGraw-Hill, pp. 79-86, 1999.
18. E. H. Nicollian and J. R. Brews, “MOS (Metal Oxide Semiconductor) Physics and Technology”, Bell Telephone Laboratories Inc., pp. 319-353, 1982.
19. D. M. Caughey and R. E. Thomas, “Carrier Mobilities in Silicon Empirically Related to Doping and Field”, *Proc. IEEE*, no. 55, pp. 2192, 1967.
20. A. Lochtefeld, I. J. Djormehri, G.Samudra and D. A. Antoniadis, “New Insights into Carrier Transport in n-MOSFETs”, *IBM J. Res. & Dev.*, Vol. 46, no. 2/3, pp. 347-357, 2002.

Chapter 4

Pulsed I-V Measurements

4.1 Introduction

In section 3.5 the procedure for the measurement of an MOSFET I_{ds}-V_{ds} characteristic was described. However, the problem of self-heating may arise during this measurement when a strained silicon device is being tested, as discussed in section 2.7. Power dissipated by an MOSFET is equal to I_{ds}^2R, where R is the resistance of the channel. Experience gathered during this investigation has shown that self-heating only begins to have a noticeable effect on I_{ds} when the dissipated power becomes quite large (>~ 0.5mWμm^{-1}); consequently it may not be apparent unless the device being measured is aggressively scaled or under high bias.

The pulsed measurement system, as discussed in section 2.7.2, is perhaps the simplest and easiest method of obtaining the "correct" I_{ds}-V_{ds} characteristic. The traditional method for performing a pulsed measurement involves a pulse generator, which must be capable of supplying pulses of the order of nanoseconds and with a very short rise time, and a digital oscilloscope capable of fast sample rates.[(1), (2)] In this chapter the development of a stand-alone system for pulsed I_{ds}-V_{ds} measurements is described. This system has possible advantages over the normal method and, once complete, does not require a fast oscilloscope. It was built in collaboration with Adrian Lovejoy of the Electronics Workshop in the Department of Physics at Warwick.

4.2 System Specification

Pulsed measurements performed on GaAs devices have encountered difficulties with noise when the pulse was applied to the gate,[3] although this has never been reported for group IV devices. A solution to this is simply to apply the pulses to the drain instead of the gate, but most authors have preferred to utilise a gate pulse, perhaps because the (relatively weak) capacitive coupling between the drain and the gate results in some fluctuation of vertical field as the drain pulse is applied. Because it was unclear whether it would be advantageous to pulse the gate or the drain, a system was constructed that was capable of applying a pulse to either the drain or the gate, or both. The timing and duration of the two pulses could be set completely independently of each other.

Originally, a personal computer fitted with a fast analogue to digital conversion card was to be in direct control of the measurements, initialising and terminating each of the pulses and digitising the resultant drain current at some specified point. A problem with the design of modern PCs soon became apparent, however. In theory the 66MHz clock speed of the PCI bus should have been capable of generating pulses as short as 15ns but it was very difficult to ensure that no other components (such as the hard drive) made demands on the system resources. If this occurred, there could be a delay of several microseconds before the computer returned to the task of pulse control. In order to circumvent this problem, the electronics constructed to create the pulses were also made capable of controlling the pulse timing and of digitising the resultant drain current. The data could then be held in a latch with a data valid flag until such a time that the PC was ready to collect it. This approach also had the benefit of allowing the software to be written in a higher programming language,

since fast response of the computer to commands was no longer critical. Capital Equipment Corporation's Testpoint™ software [4] was chosen to write the program that would control the measurement.

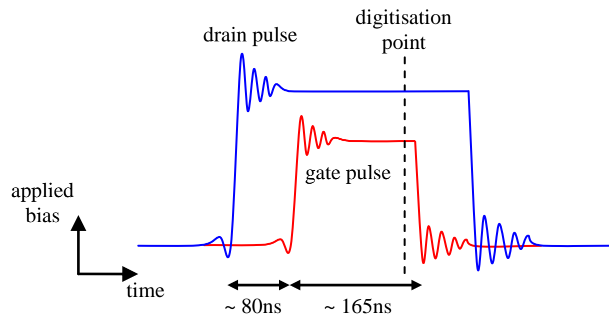


Figure 4.1. Diagram showing a typical pulse sequence during a measurement.

Digitisation of the drain current was performed towards the trailing edge of the applied pulses (figure 4.1) to allow the maximum possible settling time for capacitive and inductive effects to die away. Provision was necessary for an oscilloscope to be connected to confirm that measurement was occurring once the pulses were stable. In the case of some devices it was advantageous to pulse both the drain and the gate, for example if the threshold voltage lay the wrong side of zero, such that the device was on under no applied gate bias.

A 24MHz crystal clock controlled the timing of the two pulses, thus giving increments of 42ns. At the beginning of a measurement the computer wrote to the pulse system by means of a digital I/O card, instructing it on the polarity, magnitude and the rise and fall points of each of the two pulses. The ADLINK PCI-7200 card [5] that was used was capable of transferring 16 bit numbers, so each measurement lasted for 2.73ms ($2^{16} \times 42$ns). The digitisation point was fixed in the centre of this interval,

1.37ms (or 32768 intervals) after the measurement had commenced. The pulses ranged in magnitude from zero to 10V but could be of either polarity. Because of the way that the pulse timing was designed, control over the duty cycle was somewhat reduced. If, for example, a 1ms pulse duration was required then the duty cycle was fixed at 37% (assuming the computer commenced another measurement immediately after completion of the first). It would be simple enough to massively reduce the duty cycle by adding a software delay of, say, 0.5s to ensure that the device had cooled before applying the next pulse, but precise control of the duty cycle was not possible: it was simply kept very low in order that the results were not influenced. A further limitation of the system was that it was not possible to measure any current other than that of the drain.

4.3 System Development

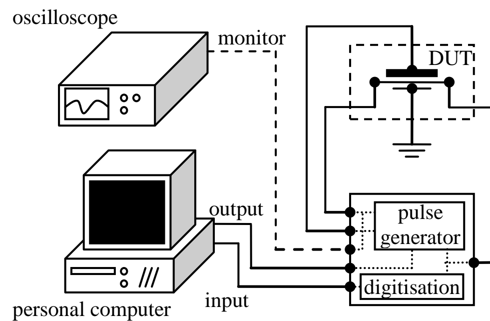


Figure 4.2. Schematic diagram of the experimental setup employed to conduct pulsed measurements of MOSFETs.

The equipment was set up according to figure 4.2. The device under test (DUT) was mounted in a shielded metal enclosure to prevent interference from external sources that could affect the validity of the measurement. In order to reduce parasitic

capacitive and inductive effects as far as possible, chips were mounted in packages with gold wire used to contact to individual devices. A substrate connection was necessary to prevent charging of the body, although this was not always a problem.

The system was calibrated by measuring known standard resistors, and then using control silicon MOSFETs, which do not suffer appreciably from self-heating. The I_{ds}-V_{ds} characteristics from the pulsed measurement system were compared with those obtained from a DC measurement using an Agilent 4156C parameter analyser, with good agreement being reached. The best results were obtained using a gate pulse, or in the case of requiring two pulses, by applying the drain pulse at least 80ns before the gate pulse. This method has advantages over the use of a constant drain bias because the duty cycle of the drain pulse was still less than 0.01%, reducing the potential for heating between measurements.

With this setup it was found that in general it was possible to use gate pulses as short as 125ns (three time increments) before capacitive and inductive effects rendered the measurement inaccurate, although under low bias conditions or with long devices this had to be increased to 200ns due to the reduced drain current. In these cases however, minimal self-heating was encountered so this was not of great concern. Successful pulsed measurements of resistors required a pulse duration of at least 83ns (two time increments), indicating that the RC time constants of components such as the device holder and even the 50Ω coaxial transmission lines may have been limiting the measurements. When pulsed measurements were performed on strained silicon nMOSFETs provided by ST Microelectronics (see section 4.4 for device details), it was found that self-heating was greatly reduced when measuring with a gate pulse of

125ns and the duty cycle less than 0.005%. However, the continuing presence of negative output conductance under pulsed conditions indicated that self-heating had not been entirely eliminated (figure 4.3).

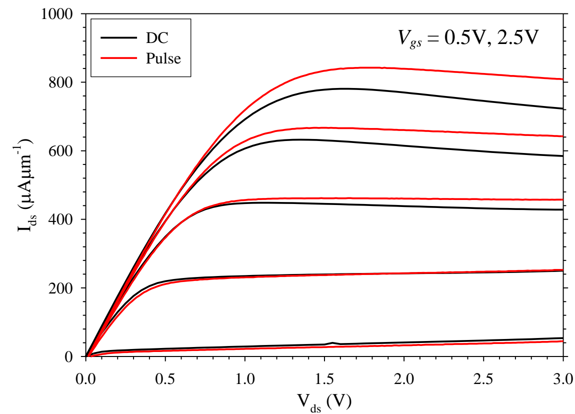


Figure 4.3. Comparison of quasi-static and pulsed (125ns) I_{ds}-V_{ds} curves for a 0.175 × 10µm strained silicon nMOSFET.

In order to reduce the pulse duration further, a modification to the existing system was constructed to address the issue of the charging spikes associated with the parasitic capacitances. Two pieces of additional equipment were required: a fast pulse generator, capable of generating pulses shorter than the 42ns limit of the original system, and an integrator. This would integrate the drain current resulting from a pulse with respect to time, so that the charging spikes observed would cancel and the current in the absence of capacitive and inductive effects would be obtained.

The leading edge of the drain pulse was used to start a monostable signal in the fast unit. The short pulse duration was set by the time taken for this signal to traverse a 50Ω coaxial transmission line; hence it could be varied by altering the length of this line. It was ensured that the output impedance of the pulse generator was low

compared to the DUT, in order that the vast majority of the potential was applied to the device.

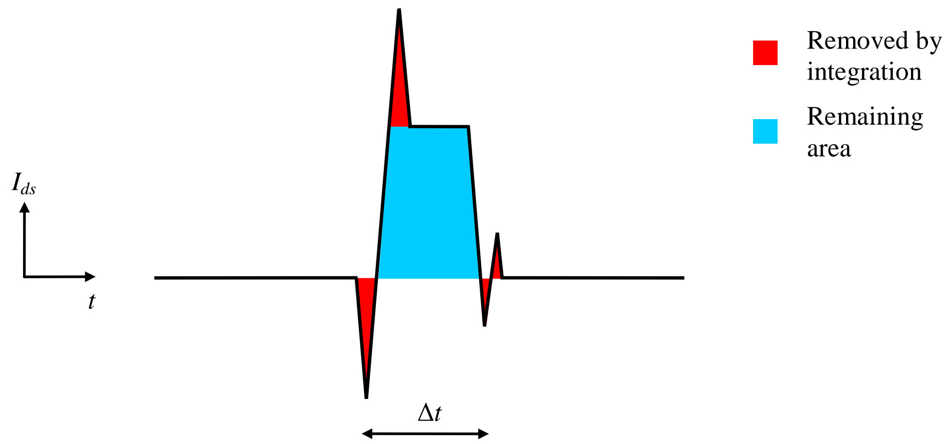


Figure 4.4. Diagram showing how the integrator was able to remove much of the uncertainty surrounding the drain current.

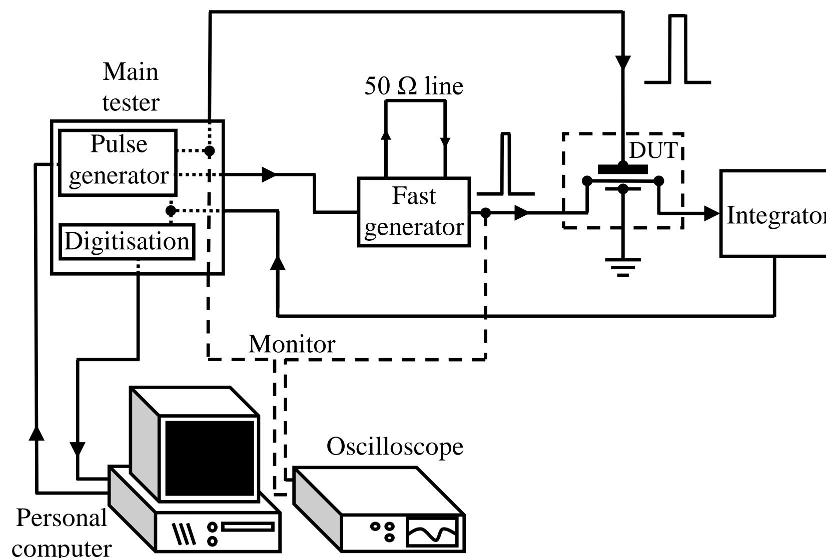


Figure 4.5. Schematic diagram of the modified system. The integrator allowed for measurement using very short pulses by removing the charging spikes encountered.

Again the system was calibrated with known standard resistors and then silicon control devices. At first the system was only capable of supplying negative fast pulses but later an additional fast pulse generator was built for the measurement of nMOSFETs. Figure 4.6(a) shows that good agreement was achieved between pulsed and quasi-static measurements for a silicon control pMOSFET, although the pulsed system was clearly in error for small drain bias. This was found to be due to incorrect drain pulse amplitudes for -0.3V < V_{ds} < 0.3V. The problem is still under investigation, but since self-heating occurred only under moderate bias, it was not of great concern in these investigations. Noise in the measurements varied widely, and for no apparent reason. When the noise was particularly bad, the results of several measurements were averaged to give a cleaner characteristic.

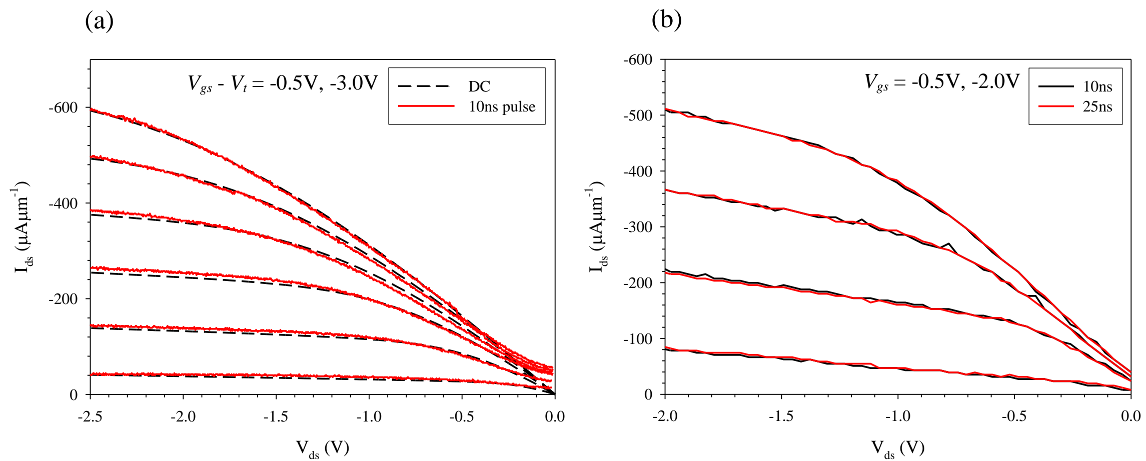


Figure 4.6. Comparison between (a) pulsed and quasi-static measurements for a silicon control and (b) different pulse durations for a strained silicon pMOSFET.

To confirm that self-heating had been removed by using a 10ns pulse, a strained silicon MOSFET of the type described in the next section was measured using 10ns and 25ns pulses. There was negligible difference between them (figure 4.6(b)) so it was evident that the pulses were now short enough.

4.4 Investigation into Self-Heating in Strained Silicon MOSFETs

4.4.1 Room Temperature Measurements

The devices used for this investigation were supplied by ST Microelectronics, Crolles, France. Fabrication of the devices and basic electrical characterisation of the devices is described elsewhere.[6] The virtual substrate was linearly graded over 2µm with a CMP step, terminating in a 40nm constant composition buffer of $Si_{0.82}Ge_{0.18}$. The strained silicon cap layer was 15nm thick and the gate oxide thickness was 3nm.

Both n and p strained silicon MOSFETs were measured using 10ns pulses and the drain characteristics were compared to the static measurements obtained using the Agilent 4156C parameter analyser (figure 4.7). Because of the higher mobility of electrons in strained silicon compared to holes (and because the nMOSFET tested was slightly shorter than the pMOSFET), the power dissipated ($I_{ds}V_{ds}$) by the nMOSFET was approximately double that of the pMOSFET for given bias conditions. Consequently the observed self-heating in this device was greater.

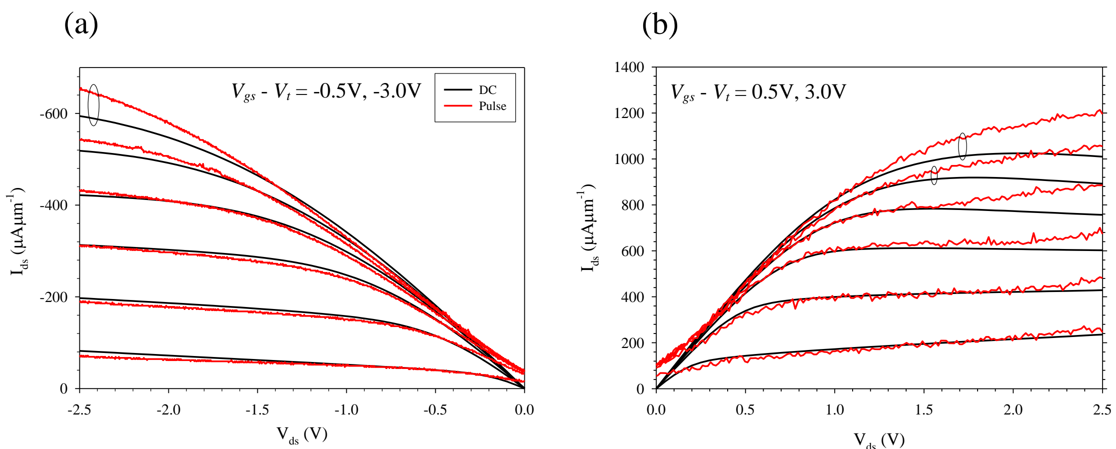


Figure 4.7. Comparison of quasi-static and pulsed measurements for a strained silicon (a) pMOSFET (L = 0.175µm), (b) nMOSFET (L = 0.15µm).

Self-heating in the pMOSFET only became noticeable when the power supply was considerably higher than a device of this geometry would typically operate with (supply voltage ~ -1.5V). This suggests that self-heating may not be a major problem facing strained silicon pMOS, at least for similar device geometries. However, in order to realise a large drive current enhancement over bulk silicon pMOS it may be necessary to use thicker virtual subtrates with a higher final germanium content, which would exacerbate the problem. The tested nMOSFET displayed much greater self-heating, consistent with the findings of Jenkins and Rim.(7) Some form of cooling solution may have to be considered if strained silicon nMOSFETs on virtual substrates are to enter mainstream production, otherwise the full performance enhancement may not be realised. This matter is discussed further in section 4.5.

4.4.2 Heated Measurements

By heating a strained silicon device above room temperature whilst performing pulsed measurements on it, it is possible to quantify the channel temperature rise that the device experiences as a result of self-heating and also to calculate the thermal resistance of the virtual substrate. This is achieved by comparison of the pulsed characteristic at elevated temperature with the DC characteristic at room temperature: the point of intersection reveals the temperature rise at a particular power dissipation.(1)

The devices described in section 4.4.1 were shielded and placed on a metal platform, which could be heated to more than 200ºC. 10ns pulsed measurements were performed at different temperatures and compared with the drain characteristic as measured by the parameter analyser (figure 4.8). From this a channel temperature rise

of 125ºC in the strained silicon pMOSFET was estimated at the highest power levels (~ 1.3mWμm^{-1}). The channel temperature exceeded 200ºC in the nMOSFET as the dissipated power increased beyond 2mWμm^{-1}.

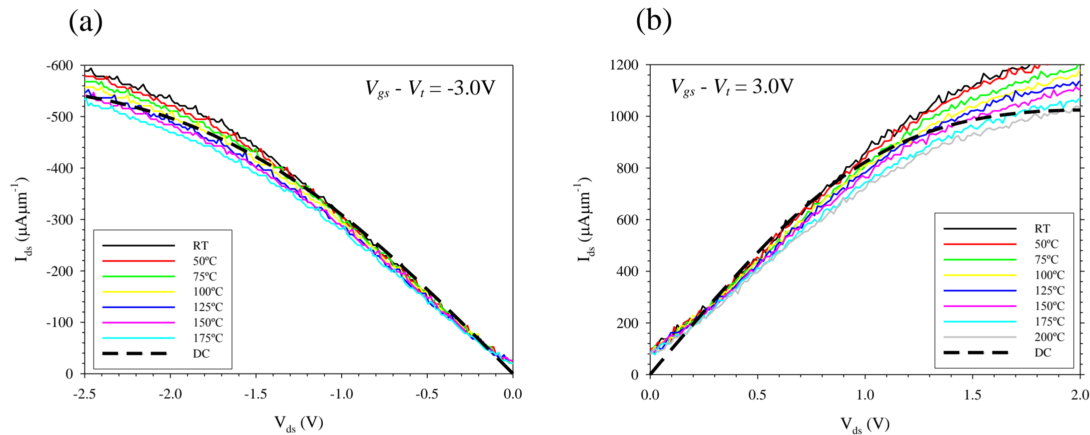


Figure 4.8. Comparison of heated pulsed and room temperature DC measurements for (a) a strained silicon pMOSFET and (b) a strained silicon nMOSFET.

The channel temperature as a function of DC power was extracted (figure 4.9) for both the nMOSFET and the pMOSFET from the crossing points of the curves in figure 4.8. At the lower temperatures, the point of intersection was not obvious due to the similar gradient of the lines being compared. The estimated errors bars in figure 4.9 are therefore decrease in size as the dissipated power increases. In each case a linear regression was performed through the data points, fixed at 22ºC for zero dissipated power. The slope of this line gives the thermal resistance of the device and is 8 ± 1KmW^{-1} for the nMOSFET and 9 ± 1KmW^{-1} for the pMOSFET.

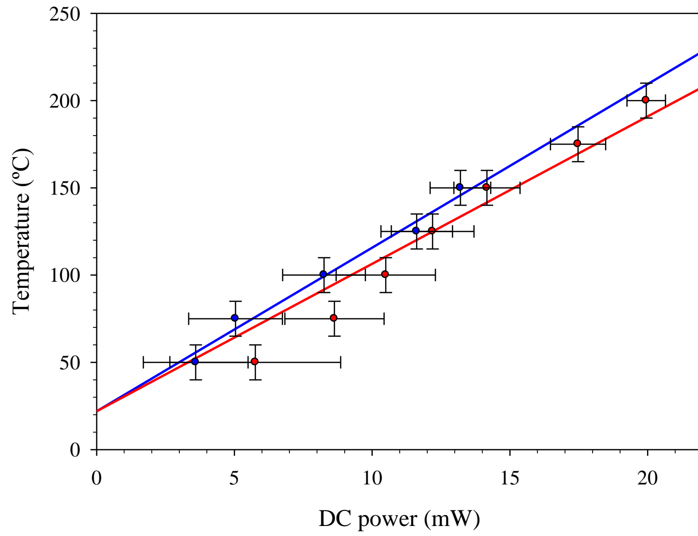


Figure 4.9. Plot of the points of intersection between the DC and pulsed curves in figure 4.8 for both the nMOSFET (red) and the pMOSFET (blue).

Jenkins and Rim adapt a model devised by Su *et al.* [8] and show that the thermal resistance of a strained silicon MOSFET may be approximated as:

$$R \approx \frac{1}{2W}\left(\frac{t_{SiGe}}{k_{SiGe}k_{Si}t_{Si}}\right)^{1/2}, \tag{4.1}$$

where t_{SiGe}, t_{Si} and k_{SiGe}, k_{Si} are the thicknesses and thermal conductivities of the SiGe and strained silicon layers respectively. Taking k_{SiGe} and k_{Si} as 0.1 and 1.5Wcm^{-1}K^{-1} respectively gives $R \approx$ 15KmW^{-1} for the devices investigated here. This is considerably larger than the values of thermal resistance extracted from the pulsed measurements and lies outside the possible range of experimental error. It should be noted that this model also somewhat overestimated the thermal resistance of the device used by Jenkins and Rim.

An assumption of the model is that the external probe pads are placed at least a *thermal healing length* away from the device to minimise their effect on internal

device cooling. The distance from the device over which the interconnect temperature falls to the substrate temperature is on the order of a thermal healing length. Jenkins and Rim ascribe the discrepancy between their result and the theoretically predicted thermal resistance to the fact that their contacts lay a distance from their device comparable to the thermal healing length.

From Su *et al.*, the thermal healing length, α, for an aluminium contact to the device is approximately given by:

$$\alpha \approx \left(\frac{k_{Al} t_{Al} t_{ox}}{k_{ox}} \right)^{1/2}, \qquad (4.2)$$

where t_{Al} is the thickness of the contacting strip and t_{ox} is the thickness of the oxide separating the strip from the semiconductor surface.

In this investigation, the contact pads were over 100µm removed from the device (figure 4.10). X-TEM was used to determine that the connecting aluminium strips were 0.5µm thick, separated from the wafer by a field oxide of 1µm thickness. This yields a thermal healing length of approximately 9µm, which is similar to the 13µm Su *et al.* calculate for their devices. It therefore seems unlikely that the discrepancy between the theoretical prediction of thermal resistance and that measured by experiment can be explained by the proximity of the contacts to the device. However, the devices investigated here had a substrate connection on top of the wafer, consisting of an aluminium contact to a heavily doped well. The aluminium was in contact with the same strained silicon layer that formed the active region of the devices. Using equation 4.2 with silicon in place of aluminium and SiGe in place of oxide yields a thermal healing length of approximately 0.7µm for the strained silicon

layer. This is much less than the distance of the source and drain contact points from the channel (~ 2.5µm) but it appears (figure 4.10) that the substrate contact lay at a distance comparable to this from the device, perhaps providing an additional path for thermal conduction that is not considered by the model. Another possibility is that significant heat conduction has taken place vertically through the field oxide to the metal of the source and drain interconnects that overlaps the heated region.

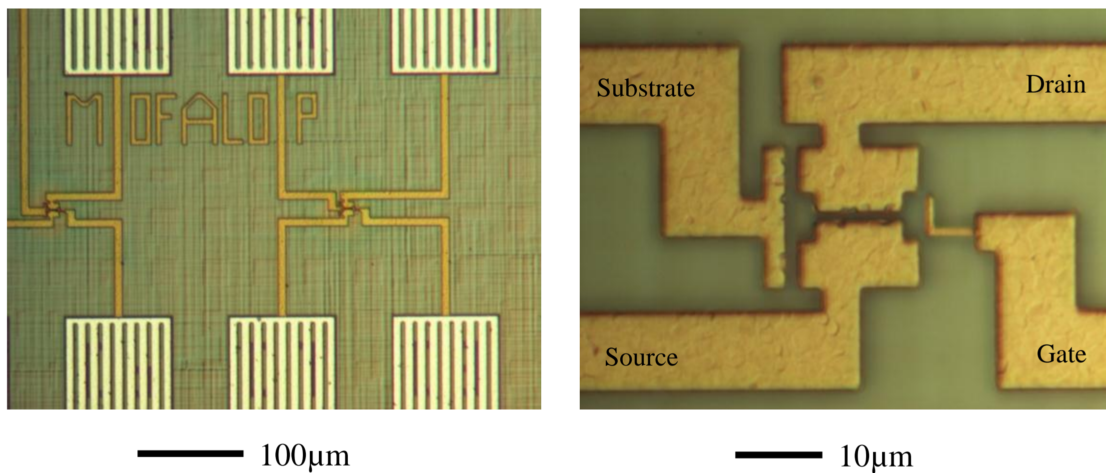


Figure 4.10. Plan view photographs of the devices used in this investigation.

4.5 Thermal Issues in the Integration of Strained Silicon

The results presented in this chapter suggest that it may be difficult to introduce strained silicon into state-of-the-art CMOS circuits because of self-heating. At elevated temperatures, strained silicon devices on virtual substrates will show a much reduced performance enhancement (if any) over standard silicon and the negative impact of elevated temperature on device reliability must also be considered. Higher temperatures accelerate both the breakdown of the gate dielectric [9] and the electromigration of metal ions in the interconnects.[10]

Using figures from the 2003 edition of the semiconductor roadmap,[11] the maximum power dissipated by a high performance nMOSFET for logic applications in the on-state is $1.33 mW\mu m^{-1}$ in 2004, with this figure rising slightly over subsequent years. A low operating power (LOP) nMOSFET dissipates a maximum of $0.48 mW\mu m^{-1}$. Using these figures, it is possible to predict how thin virtual substrates must be made in order to avoid excessive self-heating, using the adapted theory of Su *et al.*.

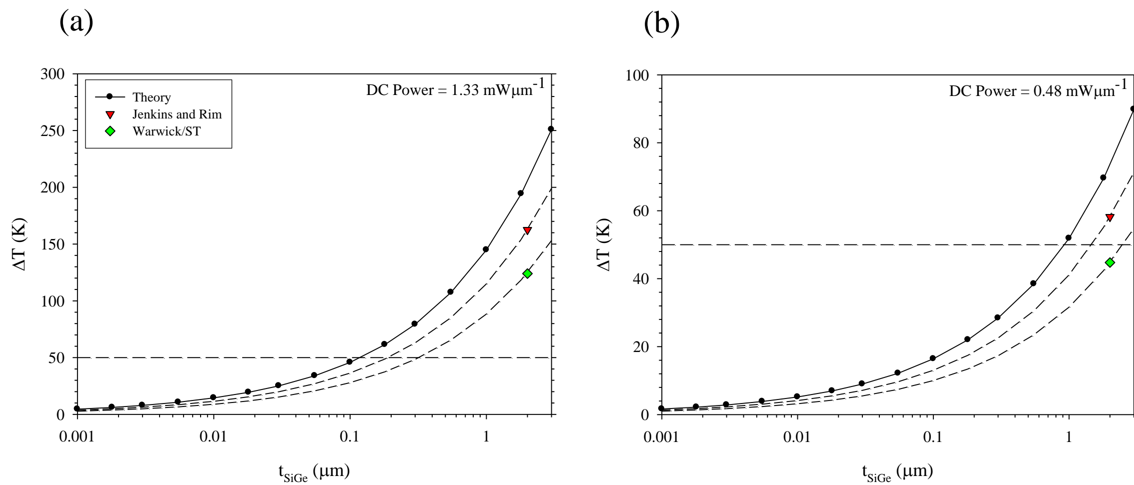


Figure 4.11. Expected temperature rise in the channel of strained silicon MOSFETs as a function of virtual substrate thickness for given DC power.

The results of Jenkins and Rim and of this work are also plotted in figure 4.11 and extrapolated assuming a constant correction factor between them and the theoretical prediction. In order to keep the temperature rise below 50K for the performance MOSFET, the virtual substrate must be of the order of 100nm thick, whilst the same constraint for the LOP MOSFET allows a virtual substrate of 1 - 2μm thickness. The figure of 50K has been chosen because these investigations have shown that a temperature rise of this magnitude only has a small effect on the saturation drain

current. The implications of a 50K rise in channel temperature for circuit reliability may be severe but this is outside the scope of this investigation.

Currently the best strained silicon device characteristics have been achieved on virtual substrates of several microns thickness,(12) which would appear to be unacceptably thick from the point of view of high performance nMOSFET self-heating. The situation is not as bad for pMOS, in which dissipated power per unit width is typically a factor of two or three lower.(13) Some progress has been made in the creation of thin virtual substrates,(14) although currently the threading dislocation densities are too high to be considered for industry and they have yet to be produced using a suitable tool for mass production.

Whilst figure 4.11 holds true for analogue applications, which may require devices to be in their on-state for considerable periods of time, the situation will not be as bad as indicated for digital circuits. In order to estimate the power dissipation that may be expected at the 90nm technology node of 2004, the work of Su *et al.* is extended here. The average power dissipation per unit active area of CMOS has historically quadrupled as transistor length has halved.(15) At the 90nm technology node, individual logic transistors have a written gate length of 53nm and it may therefore be estimated that average power dissipation in the most heavily loaded parts of the circuits (e.g. clock drivers) is 1.92mWμm^{-2}. The length of the active area for this technology is approximately 0.35μm (assuming that written gate length, contact hole width and contact spacing to the edge of the active area are all ~ 50nm), yielding a time-averaged power dissipation of 0.67mWμm^{-1}. Note that this is approximately half of the static power dissipation for this technology, as expected for the most

heavily loaded devices. Self-heating is therefore likely to be a considerable problem in some regions of a digital circuit, although it should be noted that over the entire die, average power density is a much more modest 2μWμm^{-2} (figure 4.12). This indicates that self-heating may not be problematic for the majority of devices, which will only dissipate considerable power during switching. It may be possible to make provision for additional cooling in the areas of the die that require it and thus overcome the issue of self-heating. The investigations presented in section 4.4 appear to demonstrate the feasibility of this aim.

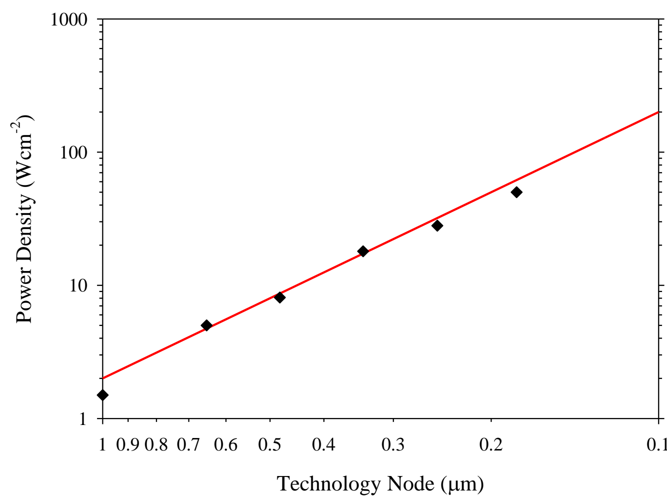


Figure 4.12. Increase in power density as transistor size has reduced (after Moore).[15]

References

1. K. A. Jenkins and J. Y.-C. Sun, “Measurement of I-V Curves of Silicon-on-Insulator (SOI) MOSFET’s without Self-Heating”, *IEEE Electron Device Lett.*, vol. 16, no. 4, pp. 145-147, 1995.
2. C. Anghel, R. Gillon and A. M. Ionescu, “Self-Heating Characterization and Extraction Method for Thermal Resistance and Capacitance in HV MOSFETs”, *IEEE Electron Device Lett.*, vol. 25, no. 3, pp. 141-143, 2004.
3. L. Selmi and B. Riccò, “Thermal Characterization of GaAs MESFETs by means of Pulsed Measurements”, *IEEE IEDM Tech. Dig.*, pp. 255, 1991.
4. Capital Equipment Corp., website http://www.cec488.com.
5. ADLINK Technology Inc., website http://www.adlinktech.com.
6. M. Jurczak *et al.*, “Study on Enhanced Performance in NMOSFETs on Strained Silicon”, *ESSDERC Conference Proceedings*, pp. 304-307, 1999.

7. K. A. Jenkins and K. Rim, "Measurement of the Effect of Self-Heating in Strained-Silicon MOSFETs", *IEEE Electron Device Lett.*, vol. 23, no. 6, pp. 360-362, 2002.
8. L. T. Su *et al.*, "Measurement and Modeling of Self-Heating in SOI NMOSFET's", *IEEE Trans. Elec. Dev.*, vol. 41, no. 1, pp. 69-75, 1994.
9. C. Hu and Q. Lu, "A Unified Gate Oxide Reliability Model", *Proceedings of the International Reliability Physics Symposium*, pp. 66-71, 1999.
10. C. -K. Hu, R. Rosenburg and K. Y. Lee, "Electromigration Path in Cu Thin-Film Lines", *Appl. Phys, Lett.*, vol. 74, no. 20, pp. 2945-2947, 1999.
11. International Technology Roadmap for Semiconductors, 2003 edition, website http://public.itrs.net/Files/2003ITRS/Home2003.htm.
12. K. Rim, J. L. Hoyt and J. F. Gibbons, "Fabrication and Analysis of Deep Submicron Strained-Si N-MOSFET's", *IEEE Trans. Elec. Dev.*, vol. 47, no. 7, pp. 1406-1415, 2000.
13. Y. Taur and T. K. Ning, "Fundamentals of Modern VLSI Devices", Cambridge University Press, pp. 175, 1998.
14. T. Hackbarth *et al.*, "High Frequency n-type MODFETs on Ultra-Thin Virtual SiGe Substrates", *Semicond. Sci. Technol.*, vol. 47, no. 7, pp. 1179-1182, 2003.
15. G. Moore, "No Exponential is Forever… but We Can Delay Forever", *International Solid State Circuits Conference*, 2003.

Chapter 5

Experimental Results

5.1 Introduction

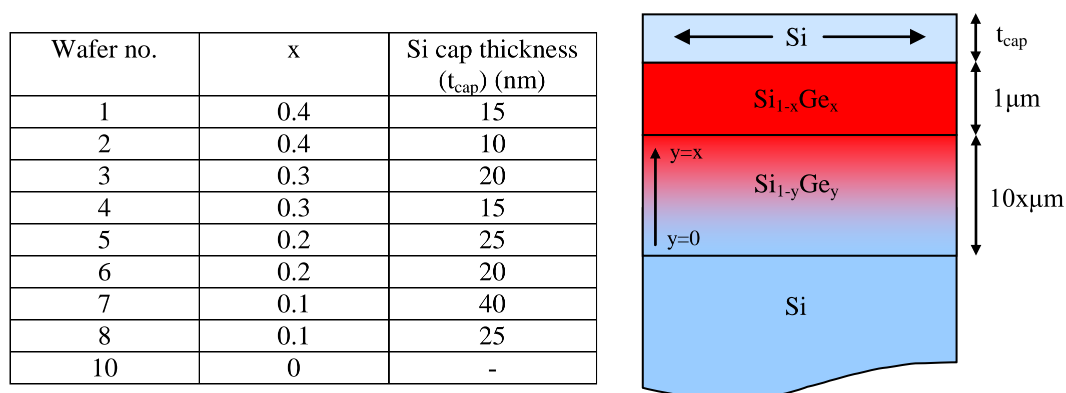
This chapter presents the experimental results arising from investigations into two batches of strained silicon wafers. The first of these (k2295) consisted of wafers grown by Dr Hans von Känel using the LEPECVD system at ETH Zürich.[1] The wafers then underwent a fast device fabrication process at Southampton University under the supervision of Dr Martin Palmer from Warwick. Characterisation of the resulting pMOSFET devices using the techniques described in chapter 3 forms the first part of this chapter.

A second batch of wafers (k2334) consisted of LEPECVD virtual substrates, again provided by Dr von Känel, although for this batch n-type wells and strained silicon layers were grown by Dr Tim Grasby using the V90S SS-MBE growth system at Warwick. The wafers underwent a standard 0.25μm process at Southampton University. Again, the resultant devices were pMOSFETs.

An additional batch of wafers was created as part of the European economic production of virtual substrates (ECOPRO) collaboration. Again, the wafers were created by SS-MBE growth of the active regions on top of LEPECVD virtual substrates. A variety of doping techniques were employed (see section 5.7) and the batch consisted of both n and p-type devices. The design of these wafers was undertaken as part of this work using a commercial device simulator. A basic electrical characterisation of some n-type devices is presented here.

5.2 Specification and Fabrication of Batch k2295

These wafers consisted of n^- silicon substrates onto which linearly graded virtual substrates were grown by LEPECVD. The grading rate was 10% germanium per micron, grown in the [001] direction. Above the graded region approximately 1µm of SiGe at constant composition was grown, on top of which lay the strained silicon cap. The background doping concentration associated with the growth chamber was approximately 5×10^{16}cm^{-3} n-type. The batch consisted of ten wafers, the specification of which is shown in figure 5.1. The silicon wafer had no epitaxial growth and was therefore not a true control.



Wafer no.	x	Si cap thickness (t_{cap}) (nm)
1	0.4	15
2	0.4	10
3	0.3	20
4	0.3	15
5	0.2	25
6	0.2	20
7	0.1	40
8	0.1	25
10	0	-

Figure 5.1. Specification of the wafers comprising batch k2295. Wafer 10 was a silicon control.

These wafers had a very thick gate oxide of approximately 100nm in an attempt to eliminate gate leakage currents. A few nanometres of thermal oxide was formed on the strained silicon in an attempt to ensure a high quality interface with a low defect and interface trap density, whilst only consuming a minimal amount of strained silicon. On top of this the remainder of the oxide was deposited by PECVD and densified by annealing in a process typically used for creation of field oxide.[2] In this

way a thick gate oxide was created without excessive consumption of the strained silicon cap. The gate material and source/drain contacts were aluminium. The maximum thermal excursion during fabrication was an RTA at 750ºC for 60s to activate the source/drain implants (50keV boron). There was no device isolation.

The mask set used for this batch of wafers included MOSFETs and capacitors. The MOSFETs ranged in length from 3μm to 300μm and in width from 25μm to 150μm, although 1000 × 1000μm devices were also provided for split C-V measurements. These devices were therefore considerably larger than the current state of the art, which was the price to be paid for the relatively swift fabrication process, since the additional steps necessary to prevent short channel phenomena such as punchthrough would have greatly increased processing time. Whilst some SCEs can be reduced using "grown in" doping profiles, this was not possible because at the time of growth the LEPECVD system used was not capable of providing such layers. Linear features were aligned to the [110] and [$1\bar{1}0$] directions of the wafer.

5.3 k2295 Structural Characterisation

Before any meaningful conclusions may be drawn from data derived from characterisation of devices, it is important to know as much as possible about the structure of those devices. This also provides some useful feedback about the success of the growth and fabrication processes in producing the devices that were originally specified. To this end, a combination of X-TEM, C-V measurements, XRD and SIMS were used with the intention of ascertaining the gate oxide thickness, strained silicon layer thickness, the final SiGe buffer composition and the material quality of the wafers.

Figures 5.2 to 5.4 show a selection of X-TEM micrographs taken in both the <004> and <220> directions; the former to reveal the thickness of individual layers and the latter to reveal the existence of dislocations as discussed in section 3.2. The threading dislocations that were present in all the strained silicon wafers are represented by the two images in figure 5.4. It is estimated that wafers 1 and 2 possessed a threading dislocation density of approximately $3\times10^{8}cm^{-2}$ and whilst this was reduced for the lower germanium content wafers, it was still disappointingly high. Figure 5.3 highlights the existence of other defects found in both the strained silicon and the SiGe.

Dr Richard Morris provided the SIMS measurements that are presented in figures 5.5 and 5.7 at the request of the author. A SIMS profile of silicon, germanium and boron was obtained for all eight of the strained silicon wafers. In addition to this, SIMS was performed on six samples taken from different points on wafer 3 in order to provide some indication of the uniformity of the wafers. The result of this investigation is shown in figure 5.6.

Quasi-static C-V measurements were used to determine the gate oxide thickness. For some of the wafers with higher germanium content it was also possible to obtain an estimate of the strained silicon thickness due to the existence of a considerable parasitic conduction path in the relaxed SiGe underlying the channel (figure 5.6).

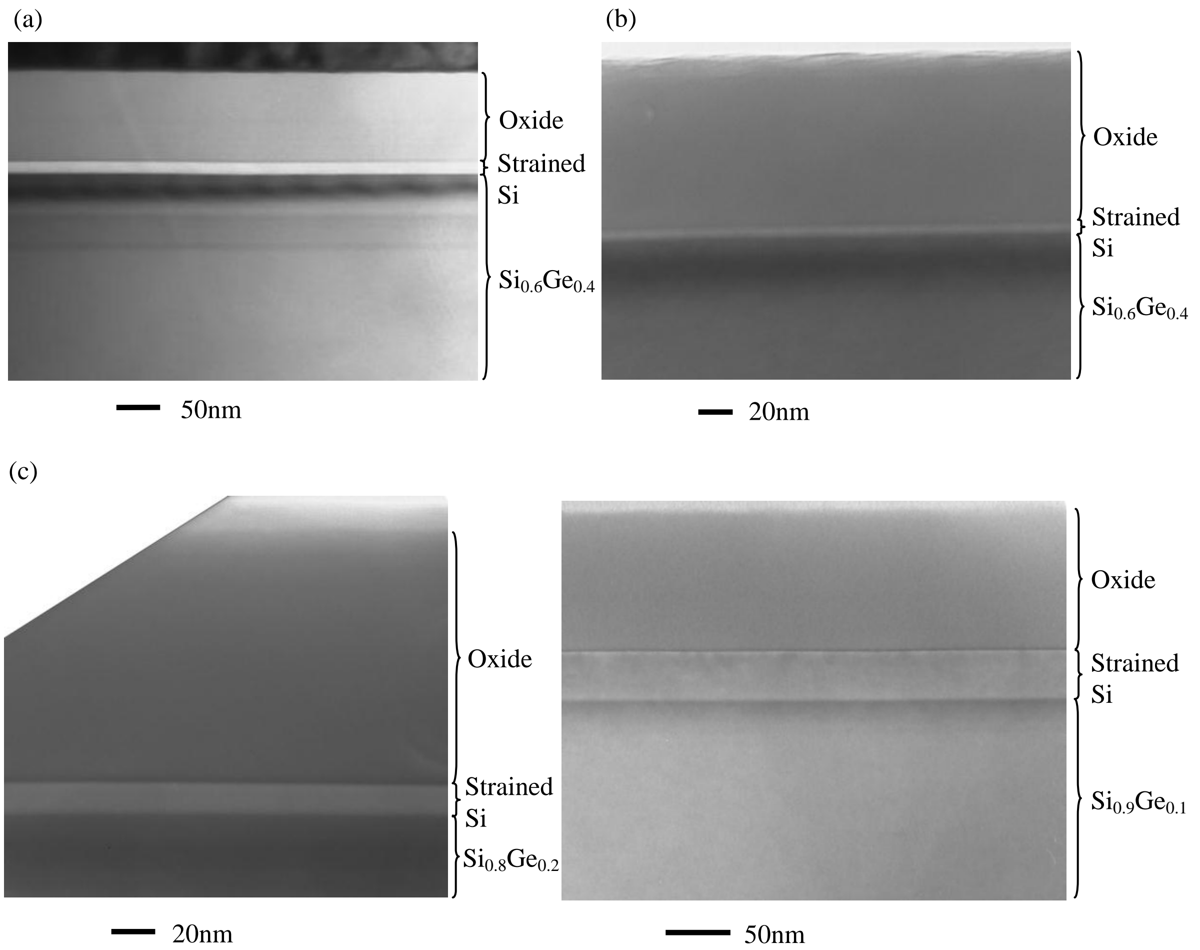


Figure 5.2. X-TEM micrographs from (a) wafer 1, (b) wafer 2, (c) wafer 6 and (d) wafer 7, taken in the <004> directions.

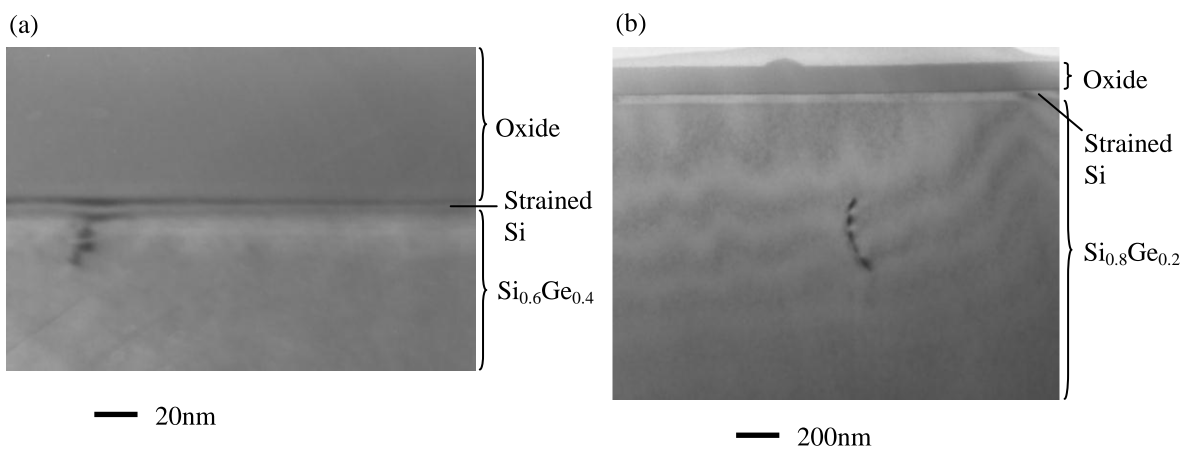


Figure 5.3. X-TEM micrographs displaying typical defects. (a) Wafer 2 and (b) wafer 6, taken in the <220> directions.

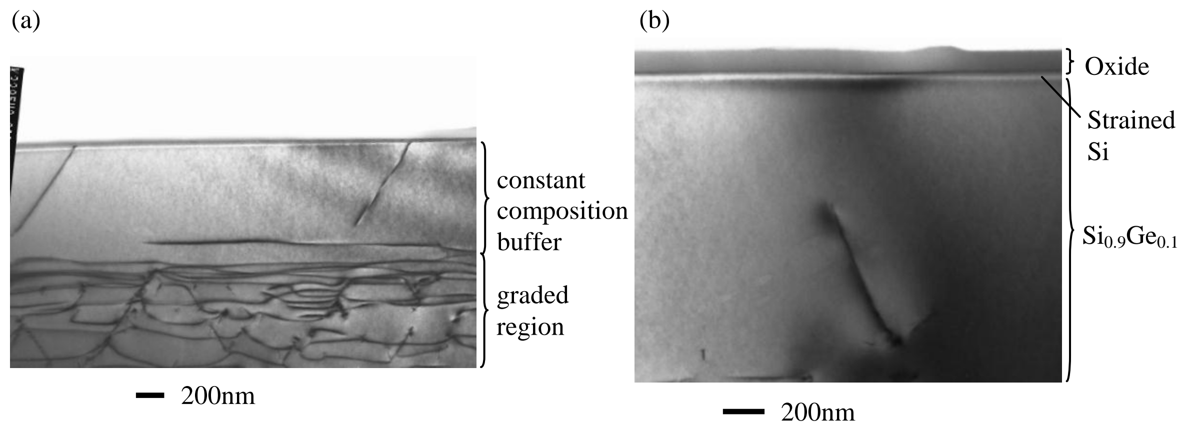


Figure 5.4. X-TEM micrographs displaying threading dislocations. (a) Wafer 2 and (b) wafer 7, taken in the <220> directions.

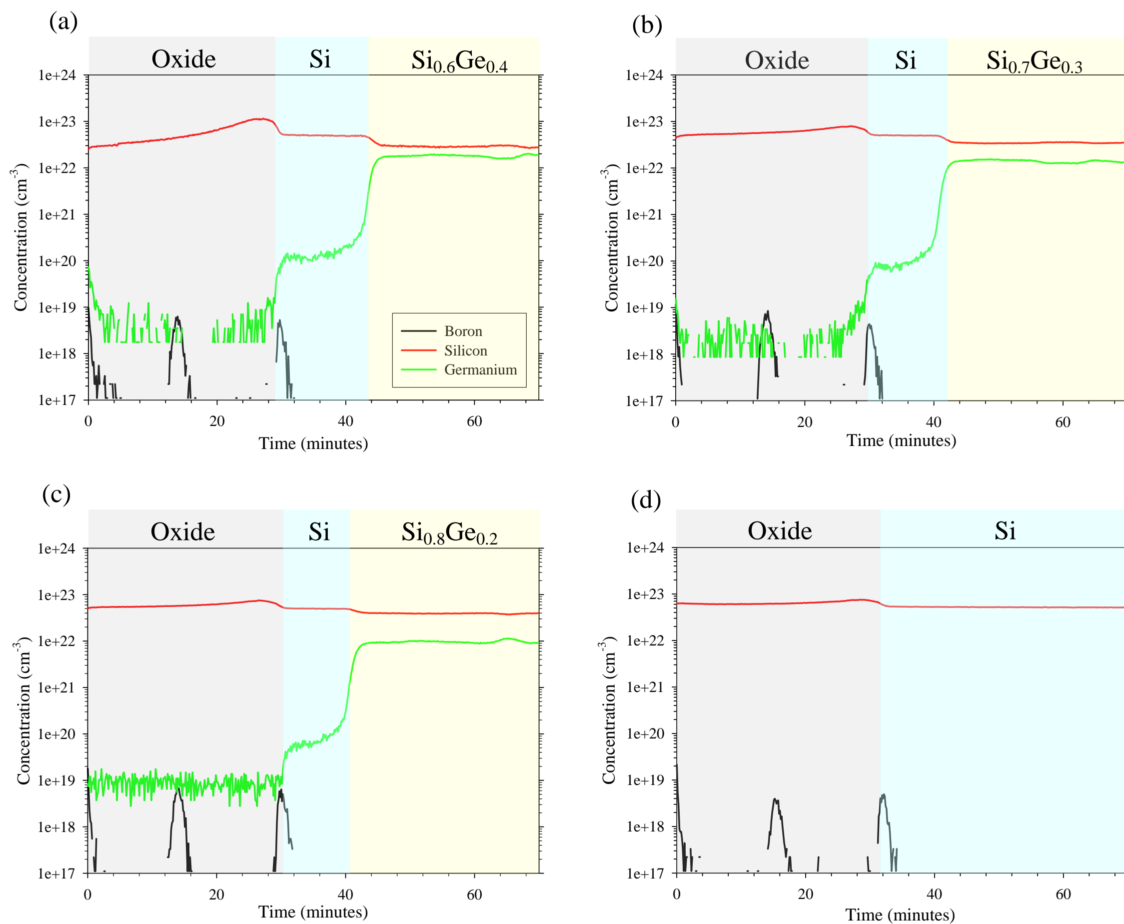


Figure 5.5. SIMS profiles of samples taken from (a) wafer 1, (b) wafer 3, (c) wafer 6 and (d) wafer 10, using O_2^+ at 500eV and normal incidence.

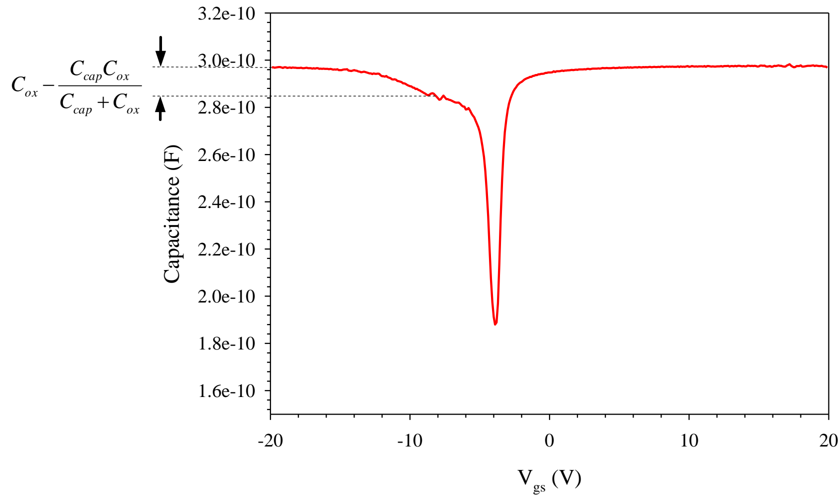


Figure 5.6. Quasi-static C-V measurement of a 1000 × 1000μm capacitor on wafer 2. The strained silicon thickness was estimated from the formation of an early plateau in inversion and results are given in table 5.1.

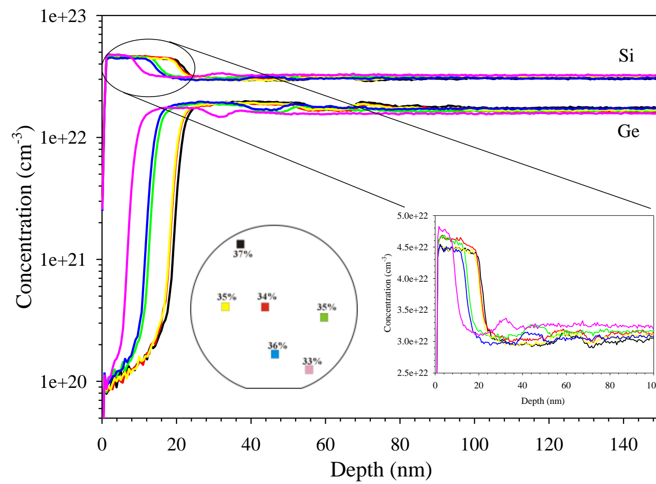


Figure 5.7. Silicon and germanium SIMS profiles of six samples taken from wafer 3, using O_2^+ at 500eV and normal incidence.

Figure 5.7 reveals that the final germanium composition varied by at least 4% across wafer 3 and also that there were fluctuations in the composition directly underneath the cap. These variations are not attributable to noise in the SIMS equipment, since noise is expected to worsen with depth. The fluctuations are only seen to a depth of

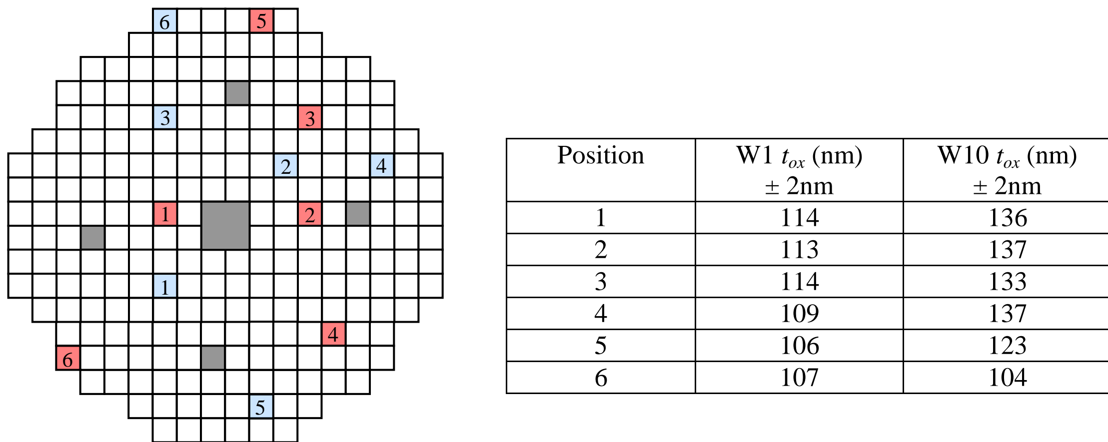
approximately 100nm, with the buffer becoming uniform below this. Additionally there appears to be a large variation in the silicon cap thickness. These inhomogeneities are possibly attributable to the fact that the wafers were not rotated during growth but held stationary. Normally there is no need to rotate the wafer when growing by CVD if the temperature and plasma density are the same across the wafer. In the regime employed by the LEPECVD system, growth rate is largely independent of temperature but variations in the plasma density could have resulted in the uneven silicon thickness. Additionally, cross-hatch on the surface of the virtual substrates may have caused small fluctuations in strained silicon thickness due to a variation in oxidation rate.[3]

Wafer	Oxide thickness (nm) ± 5nm (X-TEM) ± 2nm (C-V)	Strained Si thickness (nm) ± 2nm	Final Ge fraction ± 0.001
1: X-TEM	105	11	
XRD			0.361
C-V	116	12 ± 3	
2: X-TEM	120	6	
XRD			0.363
C-V	116	7	
3: X-TEM	122	16	
XRD			0.267
C-V	131	17 ± 4	
4: X-TEM	120	11	
XRD			0.265
C-V	130	12	
5: X-TEM	119	20	
XRD			0.176
C-V	123		
6: X-TEM	125	18	
XRD			0.175
C-V	125		
7: X-TEM	114	41	
XRD			-
C-V	140		
8: X-TEM	118	24	
XRD			-
C-V	140		
10: X-TEM	120		
XRD		-	-
C-V	139		

Table 5.1. Structural measurements of the wafers comprising batch k2295.

As can be seen from table 5.1, there are some inconsistencies between the results obtained using different methods. In particular, the oxide thicknesses as measured by X-TEM are usually lower than those calculated from C-V measurements. Rather than any inherent inaccuracy in either of these methods, this was thought to provide further evidence of the non-uniformity of the wafers. Samples for X-TEM were almost always taken away from the centre of the wafers, leaving the central areas for electrical measurements. It is conceivable that due to non-uniform heating of the wafers during oxide deposition, the oxide was thicker at the centre of wafers than at the edges. The measurements of the strained silicon layer thicknesses offer better agreement, which is somewhat surprising in light of the apparent non-uniformity highlighted in figure 5.7. It is evident that cleaning steps and thermal oxidation have removed approximately 4nm of strained silicon, assuming that these layers were as specified to begin with.

To check the uniformity of the oxide thickness across a wafer, quasi-static C-V measurements were carried out at six different points on wafer 1 (figure 5.8) and t_{ox} extracted. These findings confirm that the gate oxide varied in thickness across the wafer, being approximately 115nm at the centre and 106nm at the edges, in good agreement with the X-TEM measurements.



Position	W1 t_{ox} (nm) ± 2nm	W10 t_{ox} (nm) ± 2nm
1	114	136
2	113	137
3	114	133
4	109	137
5	106	123
6	107	104

Figure 5.8. The variation in oxide thickness measured at various points across wafers 1 (blue) and 10 (red).

It seems unlikely that the possible variation in strained silicon thickness across wafer 1 could be responsible for the oxide thickness variation since a difference of 8nm was measured between some points. Thermal oxidation comprised only a few nanometres of the gate oxide with the rest being formed by deposition, so it is likely that the variation was attributable to the latter. To confirm this and to ensure that wafer 1 was not just a “one off”, the uniformity of the silicon control was also assessed (figure 5.8) and found to display even more variation than wafer 1. The fact that the oxide is in general so much thicker on wafer 10 than on wafer 1 is something of a mystery. It has been found [(4)] that the oxidation rate of strained silicon is somewhat lower than that of bulk silicon due to the difference in bond energies, but again, since the vast majority of this gate oxide was formed by deposition, this cannot explain the difference.

XRD was performed by Dr Adam Capewell at the request of the author. It is apparent (table 5.1) that the final germanium content of the virtual substrates was slightly less than requested for wafers 1 - 6. Due to the time consuming nature of the procedure,

XRD was not performed for wafers 7 and 8, but it seems likely that the buffer of these wafers was approximately 8% germanium. The buffer layers investigated were all found to be fully relaxed. The XRD results call into question the validity of the germanium fractions determined by SIMS for wafer 3, and suggest a faulty calibration of the equipment.

5.4 k2295 Electrical Measurements

5.4.1 Doping and Interface Trap Density

The techniques described in chapter 3 were applied to discover the doping concentration in the region of the channel and the trap density at the oxide/semiconductor interface. Figure 5.9 shows a typical doping profile extracted from C-V measurements of a capacitor and selected interface trap densities as a function of energy. The midgap interface trap densities were extracted and are presented in table 5.2, together with the apparent doping concentration for each wafer. The energy of the interface traps in relation to the conduction band edge may be slightly incorrect for the strained silicon devices because of the uncertainty surrounding the band structure. The theoretical predictions of Rieger and Vogl [(5)] were used for this calculation but in any case, the important figure is the approximate trap density itself.

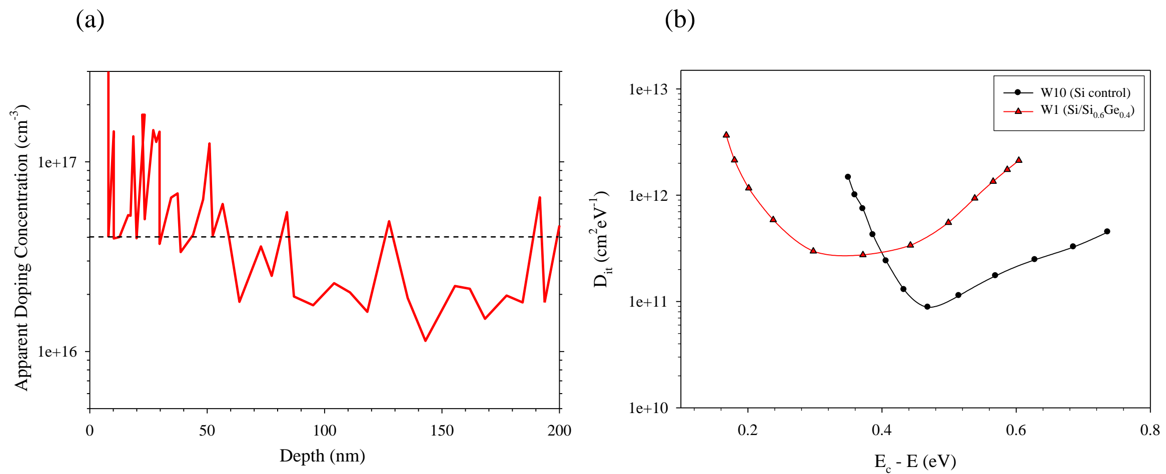


Figure 5.9. (a) Apparent doping profile extracted from a capacitor on wafer 10. These measurements were typically very noisy and serve only as an estimate of the doping concentration. (b) Interface trap density as a function of energy extracted by the high/low frequency C-V technique.

Wafer no.	Doping concentration (cm^{-3}) ± 50%	D_{it} ($cm^{-2}eV^{-1}$) ± 50%
1	1.5×10^{16}	2.8×10^{11}
2	3.0×10^{16}	3.4×10^{11}
3	3.0×10^{16}	2.8×10^{11}
4	2.0×10^{16}	4.1×10^{11}
5	2.0×10^{16}	3.3×10^{11}
6	3.0×10^{16}	3.1×10^{11}
7	2.0×10^{16}	2.7×10^{11}
8	3.0×10^{16}	2.9×10^{11}
10	4.0×10^{16}	1.3×10^{11}

Table 5.2. Extracted doping concentrations and midgap trap densities for the wafers of batch k2295.

It should be noted that owing to the inaccuracies of the methods employed to extract the values in table 5.2, the errors will be very large; possibly as much as 50%. Nevertheless, it appears that the doping concentration was in general somewhat less than expected. Even though the doping concentration was expected to be smaller than

the resolution limit of the technique, SIMS was used to check for boron, phosphorus, arsenic and antimony in addition to the usual requirement of silicon and germanium. Every wafer contained three boron spikes approaching $10^{19}cm^{-3}$; one at the oxide surface, one contained within the oxide and another at the semiconductor/oxide interface (figure 5.5). The presence of the spike within the oxide strongly suggests that this doping has been introduced at the processing stage rather than during growth, perhaps due to a contaminated furnace. SIMS reveals that these boron spikes all occur at approximately the same depth, suggesting that the oxide creation was interrupted halfway through the deposition process. The boron spikes at the semiconductor surface are the most likely explanation for the fact that the apparent doping profile extracted from the C-V measurements does not rise as sharply as normal at the surface.

The interface trap densities, whilst higher for the strained silicon wafers than the silicon control, are still indicative of a relatively high quality interface between the semiconductor and gate oxide. This is perhaps surprising in light of the SIMS evidence of considerable impurities at the interface. The fact that the strained silicon wafers had higher interface trap densities than the control is probably explained by the presence of germanium at the silicon/oxide interface. There is a tendency for germanium segregation to occur during growth, which is due to the fact that it has a lower surface free energy than silicon.(6) When a wafer with some germanium at the surface is oxidised, the interface trap density is elevated, which is thought to be because the Ge-O bond in a Si-O-Ge complex is easily broken, leaving an unpaired electron.(7) It might be expected that the wafers with a thin silicon layer (most notably wafer 2) would have a higher interface trap density because of this problem. For the

pairs of wafers at each final germanium content, those with the thinner silicon cap generally display a slightly higher D_{it}, but the difference is small compared to the possible experimental error. There is no apparent correlation between germanium content and interface trap density.

5.4.2 Series Resistance and Effective Length

The series resistance and effective channel length were now established. ΔL was expected to be relatively large for these devices, as no processing steps such as sidewall spacers had been employed to control the spread of the source/drain dopant under the gate, and the n-type doping concentration of the substrate was low. The results of these measurements are presented in table 5.3. All measurements were made on 100μm wide devices. Representative plots used for the resistance versus length and double regression methods are shown in figure 5.10.

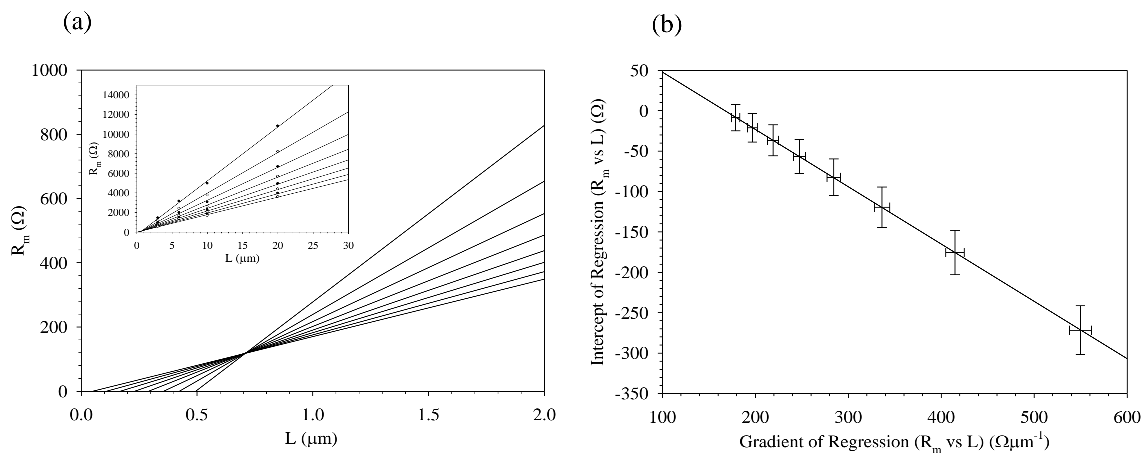


Figure 5.10. (a) Resistance versus length and (b) double regression method of extracting R_{sd} and L for devices on wafer 2.

Wafer no.	R vs L R_{sd} (Ωμm) ± 1000Ωμm ΔL (μm) ± 0.1μm	Double regression R_{sd} (Ωμm) ± 1000Ωμm ΔL (μm) ± 0.1μm
1	11500 0.73	12200 0.73
2	12000 0.72	11900 0.71
3	17000 ± 2000 0.30 ± 0.3	14000 ± 2000 0.32 ± 0.3
4	12000 0.53	11300 0.64
5	14000 0.35	14100 0.34
6	13000 ± 2000 0.90 ± 0.3	13100 ± 2000 0.96 ± 0.3
7	15000 0.65	13500 0.65
8	20000 0.85	20000 0.85
10	19500 1.05	21500 1.12

Table 5.3. Source and drain series resistance and effective length correction for the wafers of batch k2295.

The silicon control appeared to have a shorter effective length than the strained silicon wafers. This was probably due to the higher solid solubility of boron in SiGe, resulting in reduced spread of dopants under the gate stack and a reduced source/drain series resistance. In general the results of the two methods for a given wafer lay within the experimental error of each other, although those values found using the double regression method appeared to be more reliable, and it was these that were used in subsequent calculations. Because no special steps were taken to control it, R_{sd} was rather large for this batch of wafers

5.4.3 Effective Width Determination

A feature of the device batch, due to the absence of device isolation, became apparent as electrical measurements progressed. As figure 5.12 shows, when six MOSFETs of

identical gate length but differing widths were measured, it was found that the drain current per unit width was greatest for the narrowest device and least for the widest device, when it was anticipated that they would be broadly the same. The explanation for this is thought to lie in a fringing effect, whereby additional current can flow around the region defined by the source and drain as shown in figure 5.11. This effectively adds an additional width, ΔW, to the written device width. From figure 5.12(b), a value of $\Delta W = 11\mu m$ was found for wafer 1. All the wafers were found to have $\Delta W \approx 10\mu m$.

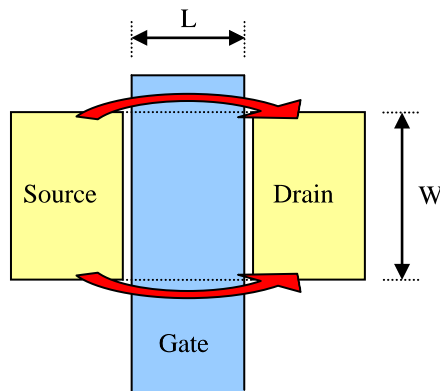


Figure 5.11. Plan view of a device, showing the path that "fringing" current can take.

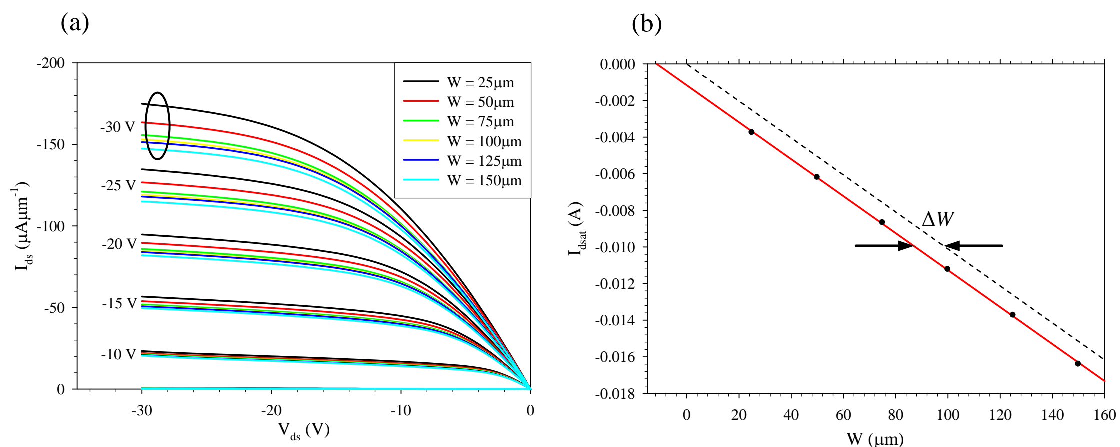


Figure 5.12. (a) Variation of drain current with device width for $L = 6\mu m$ pMOSFETs on wafer 10. (b) Extraction of ΔW for pMOSFETs on wafer 1.

5.4.4 Mobility

Using the 1000 × 1000µm pMOSFETs, the effective mobility was obtained as a function of effective vertical field from a combination of I_{ds}-V_{gs} and split C-V measurements (figures 5.13 and 5.14), and is displayed in figure 5.15. In general, noise was somewhat reduced when the C-V measurements were made outside of normal office hours.

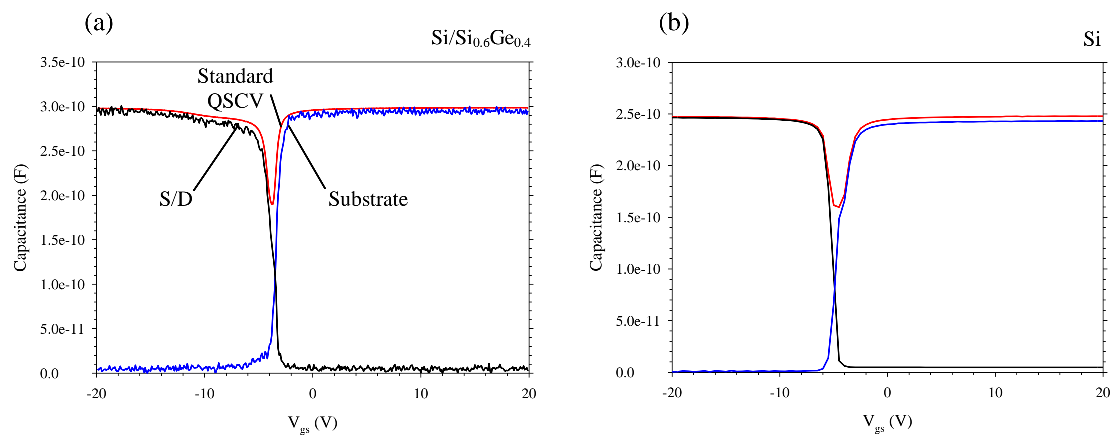


Figure 5.13. Example split C-V and standard quasi-static C-V measurements used in mobility extraction for 1000 × 1000µm devices on (a) wafer 1 and (b) wafer 10.

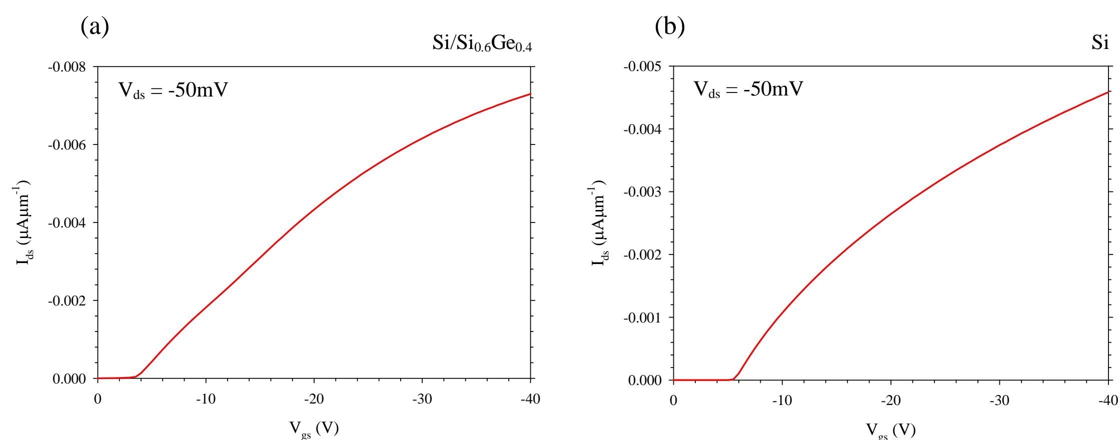


Figure 5.14. Example I_{ds}-V_{gs} characteristics used in mobility extraction for 1000 × 1000µm pMOSFETs on (a) wafer 1 and (b) wafer 10.

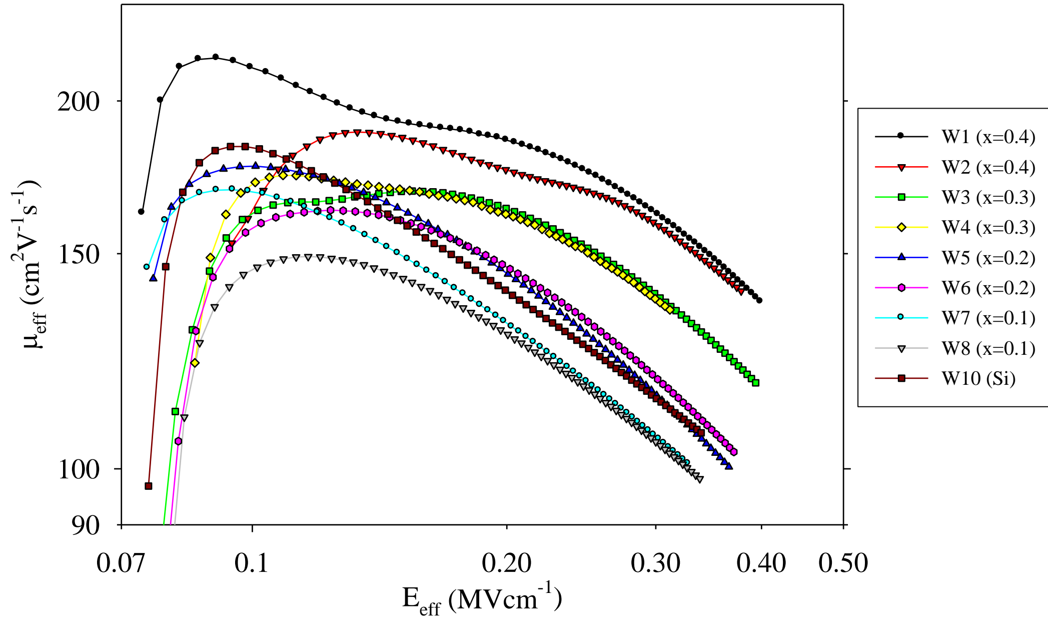


Figure 5.15. Dependence of mobility on vertical effective field for 1000 × 1000μm pMOSFETs (all wafers).

For the higher percentage germanium wafers there were two distinct peaks in the mobility, which confirmed the existence of parasitic conduction. For wafer 2, significant population of the thin strained silicon layer did not occur until the vertical field approached $0.3MVcm^{-1}$. It is these higher effective fields that are of most interest, since they allow a comparison between the carrier properties of strained and unstrained silicon. It is apparent that the largest mobility increase occurred for the largest degree of strain applied, and the mobility enhancement factor (μ_{SS}/μ_{Si}) in this case is approximately 1.4 at moderate field ($0.35MVcm^{-1}$). Wafers 3 and 4, the 30% germanium virtual substrates, showed a mobility enhancement factor over silicon of approximately 1.2. However, wafers 5 and 6 showed little or no enhancement over silicon and the mobility appeared to be following a different field dependence to that of silicon. Wafers 7 and 8, strained silicon on a 10% germanium buffer, showed lower mobilities than the silicon control. The poor material quality, revealed in

section 5.3, together with parasitic conduction probably explains the fact that all the enhancements were lower than expected. Cross-hatch was also visible on many of the wafers. It should be noted that there was no way of directly comparing the low field mobility improvements to the predictions of Oberhüber *et al.*.[8] Whilst it was desirable to probe higher vertical fields than those presented here, in general it was not possible to apply more than -40V to the gate of the large FETs before breakdown occurred. In an effort to discover something about the interface roughness of the wafers, I_{ds}-V_{gs} measurements on smaller devices were used. This was because the smaller devices could often withstand a larger field, perhaps due to the reduced probability of the gate oxide containing a serious defect, resulting in early breakdown.

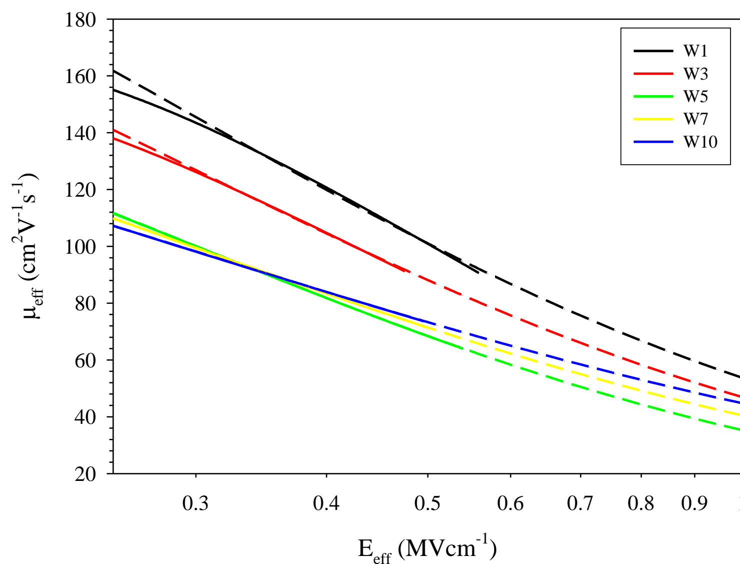


Figure 5.16. Mobility dependence on electric field for 100 × 100μm pMOSFETs. Dashed lines represent empirical fits to the data.

Accurate C-V data was sometimes difficult to obtain for the 100 × 100μm devices, owing to noise. Therefore, in order to generate the effective field data displayed in figure 5.16, split C-V measurements of the larger FETs on the same chips were used.

It was confirmed that the threshold voltages of the two devices were very similar and therefore for a given V_{gs} the vertical field should also be very similar. As it was impossible to make C-V measurements past ~ -40V, the effective field for gate voltages up to -60V was calculated by assuming a linear dependence of carrier sheet density on applied gate voltage.

At room temperature, Takagi *et al.* [9] find that there is no single dependence of hole mobility on electric field of the form $\mu_{eff} \propto E_{eff}^{\gamma}$. Curves were therefore fitted to the data in the highest accessible field region (≥ 0.3MVcm^{-1}) using the universal mobility relation proposed by Watt and Plummer:[10]

$$\mu_{eff} = \frac{\mu_0}{\left(1 + \left(E_{eff} / a \right)^{b}\right)} . \quad (5.1)$$

At high fields the population of the SiGe becomes less significant and the mobility is partially dependent on roughness at the Si/SiO_2 interface. For holes in silicon, a = 0.27MVcm^{-1}, b = 1 and μ_0 = 240cm^{2}V^{-1}s^{-1}. At first the computer was given free reign to find the values of μ_0, a and b which gave the best fits to the data (table 5.4).

Wafer no.	μ_0 (cm^{2}V^{-1}s^{-1})	a (MVcm^{-1})	b
1	196	0.516	1.85
3	190	0.454	1.66
5	198	0.304	1.29
7	212	0.268	1.08
10	197	0.298	1.01

Table 5.4. Parameters generated for a selection of wafers by fitting to the experimental data using Watt's mobility relation.

The silicon control device performed largely as expected, with the parameter *b* being very close to unity, although the low field mobility was somewhat less than normal. For the strained silicon devices it was found that that the mobility dependence on electric field became stronger as the germanium content increased, implying that the interface became rougher with increasing strain. However, the parameter *a* also increased, suggesting that a larger vertical field was required for the strained silicon devices before mobility degradation due to interface roughness began to take effect. Of course, it is important to note that mobility degradation at high vertical fields is unlikely to be solely due to increased scattering at the Si/SiO_2 interface. It is thought that the effect of quantum confinement is to cause a splitting of the light and heavy hole bands, however this separation is opposite in direction to that induced by tensile strain.(11) Consequently, at the moderately high fields investigated here, the strained silicon devices may be suffering from increasing phonon-assisted interband scattering, as the energy separation of the heavy and light hole bands is decreased.

Because it was not clear why the strained silicon devices should exhibit larger values of parameter *a* than the silicon control, curves were again fitted with this parameter fixed at $0.298MVcm^{-1}$, as extracted for the control (table 5.4). The results of these regressions are displayed in table 5.5 and are plotted as dashed lines in figure 5.16.

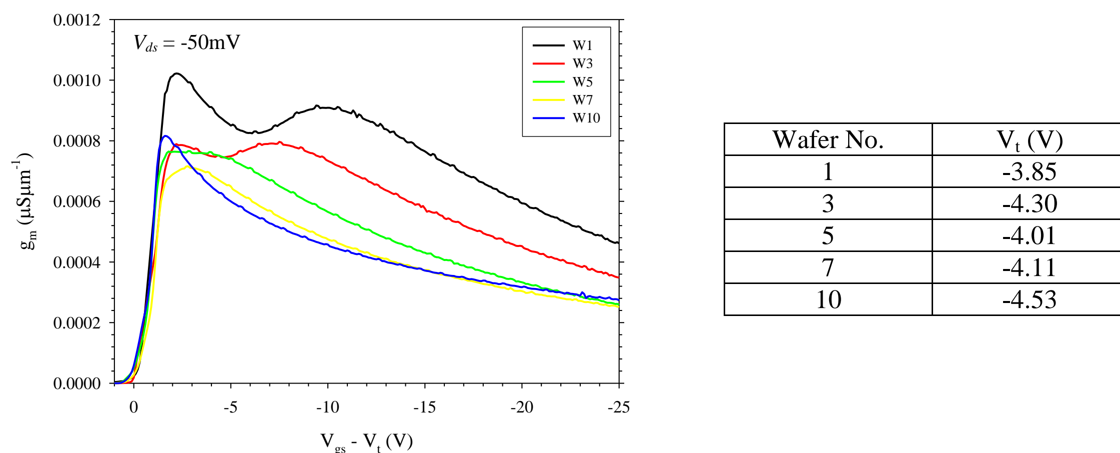
Wafer no.	μ_0 ($cm^2V^{-1}s^{-1}$)	*b*
1	292	1.23
3	255	1.23
5	201	1.28
7	200	1.13
10	197	1.01

Table 5.5. Parameters generated by fitting to the mobility data with the value of *a* fixed.

Although the agreement with the experimental data was not quite as good when using the values in table 5.5 in place of those in table 5.4, the low field mobilities generated using this method appear much more reasonable, as they vary monotonically with strain. A low field mobility enhancement factor of almost 1.5 for wafer 1 over the silicon control is indicated. This is considerably less than the enhancement factor of 2.2 predicted by Oberhüber *et al.* for strained silicon on a relaxed buffer of 40% germanium. As may be seen from figure 5.16, the mobility enhancements offered by strained silicon are predicted to be greatly reduced as the vertical effective field approaches 1.0MVcm^{-1}. In reality the enhancement (if any) for this field is likely to be even smaller than indicated for wafers 1 and 3, as the fitted mobility does not follow the experimental data for these wafers exactly.

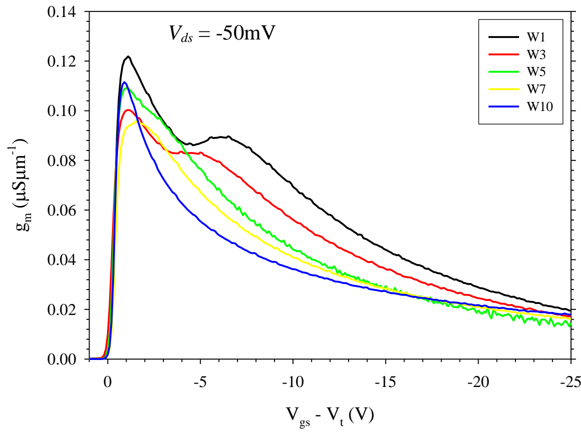
5.4.5 Transconductance and Subthreshold Swing

Transconductance as a function of gate overdrive for devices on a selection of wafers is shown in figures 5.17 to 5.20, together with some extracted threshold voltages.



Wafer No.	V_t (V)
1	-3.85
3	-4.30
5	-4.01
7	-4.11
10	-4.53

Figure 5.17. Linear transconductance as a function of gate overdrive for 300 × 100µm devices, together with calculated threshold voltages.



Wafer No.	V_t (V)
1	-3.48
3	-4.01
5	-4.23
7	-4.23
10	-4.32

Figure 5.18. Linear transconductance as a function of gate overdrive for 3 × 100µm devices, together with calculated threshold voltages.

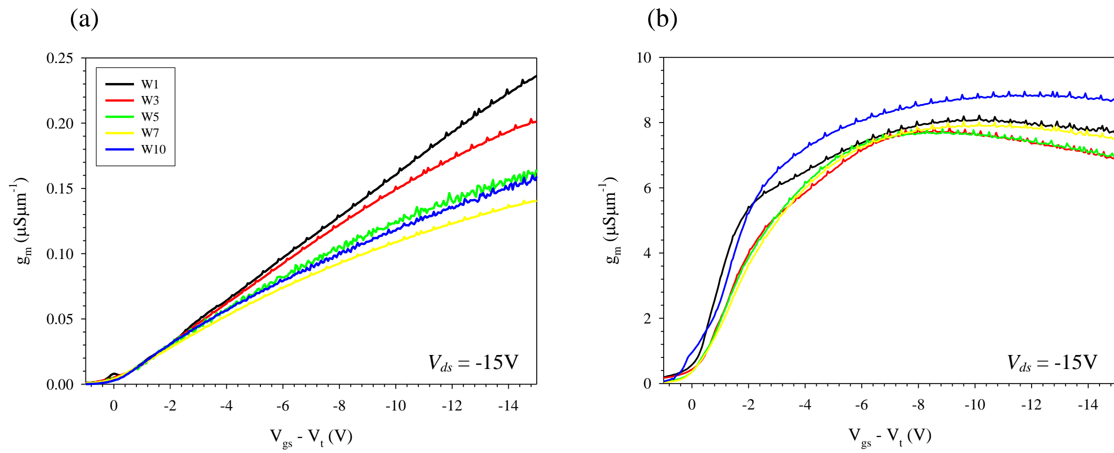


Figure 5.19. Saturation transconductance as a function of gate overdrive for (a) 300 × 100µm and (b) 3 × 100µm devices.

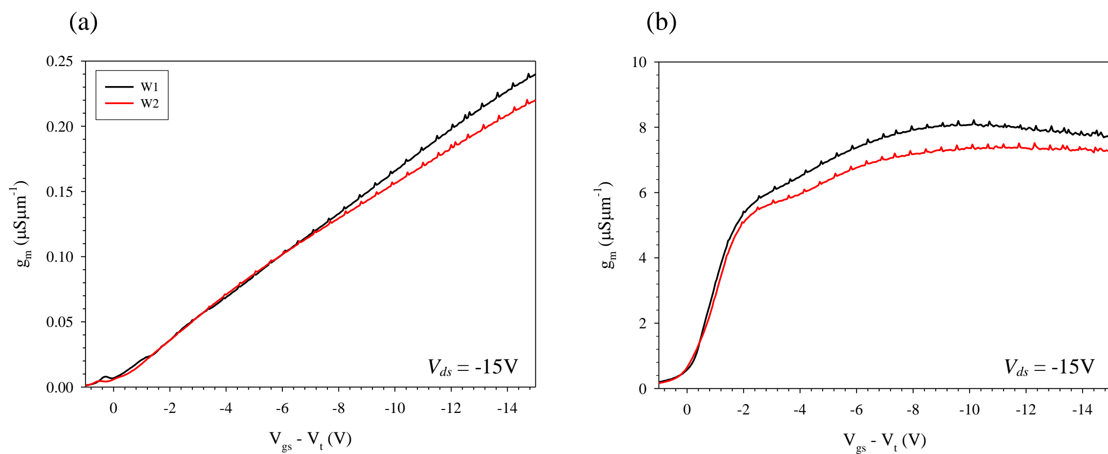


Figure 5.20. Comparison of saturation transconductance for (a) 300 × 100µm and (b) 3 × 100µm devices on wafers 1 and 2.

The transconductance measurements taken in the linear region of MOSFET operation, with low V_{ds}, further emphasize the existence of substantial parasitic conduction in the relaxed SiGe buffer. The second peak, clearly visible in the data from wafers 1 and 3 (figures 5.17 and 5.18), corresponds to population of the strained silicon layer and provides these devices with a substantial transconductance enhancement across the gate overdrive range. For the long device on wafer 1, the first transconductance peak, corresponding to population of the SiGe buffer, is higher than that of the control. This suggests that the mobility of relaxed SiGe may be higher than that of silicon if the germanium content is high enough. Care must be taken in interpreting these results however, because the control device almost certainly had a thicker oxide than some of the strained silicon devices, causing a reduction in drain current and consequently in transconductance. The mobility of holes in relaxed SiGe has been found to be similar to that of relaxed silicon,(12) although in these structures a small enhancement may be apparent due to the fact that the SiGe buffer is effectively a buried channel and may be slightly compressively strained at the Si/SiGe interface.

The measurements of transconductance in saturation on the long devices (figure 5.19(a)) also show wafers 1 and 3 outperforming the silicon control. This is a more relevant result from the point of view of devices as they would be utilised. Recalling from section 2.2 that the drain current for a long MOSFET is given by:

$$I_{ds} = \mu \frac{W}{L} C_{ox} \left[(V_{gs} - V_t) V_{ds} - \frac{V_{ds}^2}{2} \right], \qquad (5.2)$$

the saturation current, I_{dsat}, occurs for $V_{ds} = V_{gs} - V_t$. Substituting into equation 5.2 and differentiating with respect to V_{gs} gives the expression for saturation transconductance in a long device:

$$g_{msat} = \mu \frac{W}{L} C_{ox} \left(V_{gs} - V_t \right). \quad (5.3)$$

In the absence of mobility degradation, the saturation transconductance is therefore linearly dependent on the applied gate bias. In reality, mobility is reduced as the gate bias increases due to interface roughness scattering, explaining the shape of the graphs in figure 5.19(a) and figure 5.20(a). The device on wafer 1 suffers the least mobility degradation and is therefore closest to a linear relationship between g_{msat} and V_{gs}. At the point where V_{gs} - V_t exceeds V_{ds}, the pinch-off region vanishes completely and the device leaves the saturation region of operation, with the result that the transconductance begins to fall due to mobility degradation.

It is evident that the strained silicon devices did not perform as well at a written channel length of 3μm. In the linear regime, strained silicon still offers a performance enhancement over the control, although it is reduced in comparison to that at 300μm. Interestingly, the strained silicon also appears to populate under a smaller gate overdrive than is the case for long channels, as may be seen from the shift in the position of the second peaks. Given that the 3μm devices were measured on the same chips as the 300μm, this is unlikely to be due to variations in oxide or strained silicon thickness. The explanation may lie instead with device processing. There is likely to be a negative effect on carrier transport at the ends of the channel in proximity to the source and drain implants due to increased ionised impurities. It is conceivable that this effect was increased in the strained silicon devices, due to the increased solid solubility for boron exhibited by SiGe. This would lead to a large concentration of doping at the SiGe surface, at the same depth that the conduction holes in this layer occupy. Any negative effects associated with the ends of the channel would therefore be increased for the strained silicon devices. The existence of elevated boron level in

this region could also have the effect of bringing about an earlier population of the strained silicon layer and of reducing its mobility, due to damage. If this hypothesis is correct then we would expect the wafers with a lower germanium content and a thicker strained silicon channel to be affected less relative to the control as the channel length is reduced. A comparison of figures 5.17 and 5.18 appears to confirm this prediction. It should be noted that for this batch of wafers, no effort was made to confine the source/drain doping to a tightly defined region. In state of the art devices, sidewall spacers are routinely used for LDD formation, and also control the lateral movement of dopants under the gate stack.

Measurements of saturation transconductance for the 3µm devices (figure 5.19(b)) revealed that the silicon control outperformed all of the strained silicon devices. In this instance it is likely that velocity saturation is occurring, explaining the loss of the linear dependence of transconductance on gate overdrive seen for the long devices. Carriers in strained silicon are not thought to have a saturation velocity any higher than in unstrained silicon, so much of the enhancement due to mobility is lost. The saturation velocity of holes in SiGe is known to be lower than that of silicon,[13] so the higher percentage wafers are likely to be suffering in this regard as well. Furthermore, significant self-heating was occurring, particularly for wafers 1 and 3, with the thickest virtual substrates and the greatest drain currents (section 5.4.7).

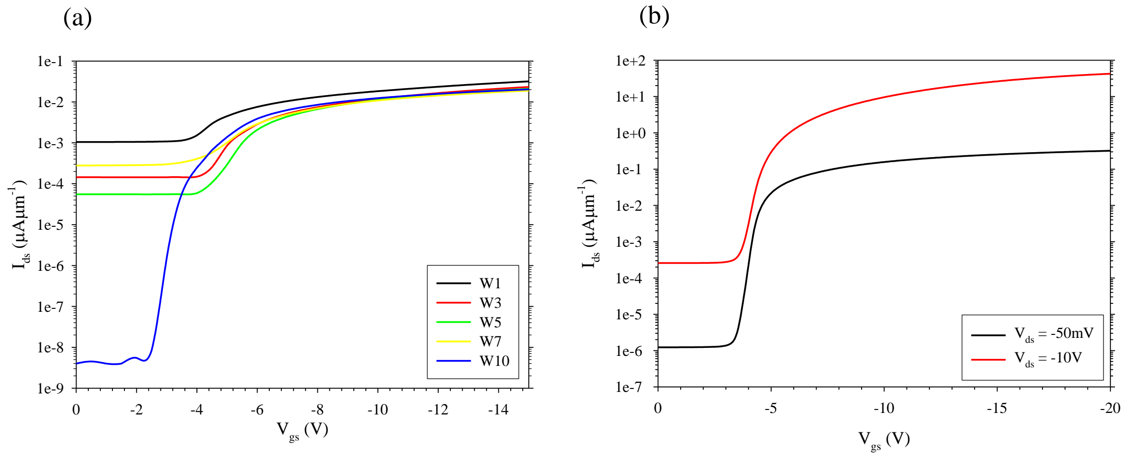


Figure 5.21. (a) Turn-on characteristics of pMOSFETs on various wafers (L = 100µm, V_{ds} = -50mV). (b) Turn-on characteristics of a good device on wafer 5.

Owing to the high threading dislocation density uncovered by X-TEM, I_{off} for all of the devices on the strained silicon wafers is extremely poor (figure 5.21(a)). In the case of wafer 1 there was often less than two decades of drain current variation between I_{on} and I_{off}. It was confirmed that this was not due to gate leakage, as gate currents were many orders of magnitude smaller than body currents. The silicon control had an I_{on}/I_{off} ratio of 10^6, which is also relatively poor considering the length of the devices measured here. The subthreshold swing for devices on wafer 10 was approximately 200mV/decade, but no extraction was possible for many of the strained silicon devices. Some of the smaller devices on wafers 5 - 8 did have reduced off-currents, probably due to a lower than typical threading dislocation density in those regions. For a particularly good 10 × 50µm device on wafer 5 which showed five decades of variation in I_{ds}, the subthreshold swing was extracted as 211mV/decade (figure 5.21(b)). At first sight this seems very poor but is in fact expected for this geometry, owing to the thick gate oxide. Equation 3.8 predicts a subthreshold swing of 185mV/decade for a surface channel device with a similar doping concentration

and oxide thickness, which rises to 188mV/decade when the capacitance of the strained silicon cap is taken into account. The slight discrepancy between theoretical and measured values can probably be attributed to interface traps.

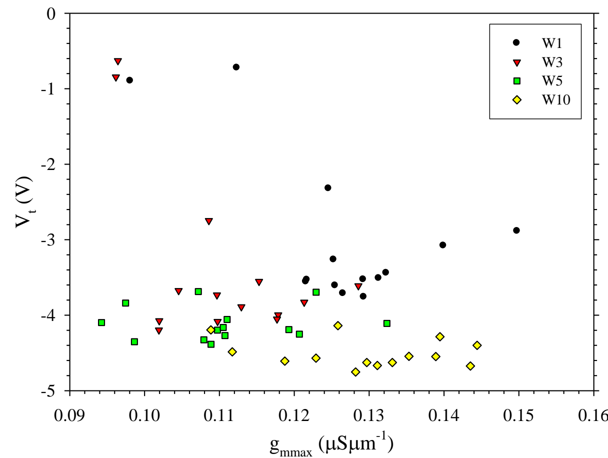


Figure 5.22. Scatter plot of peak transconductance and threshold voltage for

3 × 150μm pMOSFETs with V_{ds} = -50mV.

Confirmation of the non-uniformity of the batch is provided in figure 5.22, which was generated from measurements across the wafers. This demonstrates that all wafers, including the silicon control, had a large spread in peak transconductance. In the case of wafer 10 the best device showed a 30% improvement over the worst and the spread became slightly larger for the strained silicon wafers. This is probably related to the variation in gate oxide thickness across the wafers, since devices with a thicker oxide will naturally have lower transconductance. The threshold voltages of the silicon control devices were at least reasonably consistent, which was not the case for the strained silicon devices on the higher germanium percentage substrates. This is likely to be attributable to variations in the thickness of the strained silicon layer, although some devices turned on at much smaller V_{gs} than the majority, which may be the result of a particularly high boron concentration in those areas.

5.4.6 Saturation Drain Current

The drive current of a MOSFET is an important parameter in determining CMOS circuit performance because the drain output of one device is routinely used to charge the gate of another.[14] I_{ds}-V_{ds} curves were measured for long devices from all wafers, thus avoiding the problem of self-heating (see section 5.4.7). Figure 5.23 shows that the device from wafer 1 displayed approximately 60% increase in I_{dsat} over the silicon control, whilst the wafer 3 device displayed ~ 30% improvement. The results displayed are typical for devices of L = 300μm but some variation was seen due to the non-uniformity of the wafers. Again, the variation in gate oxide thickness will have affected the drain current.

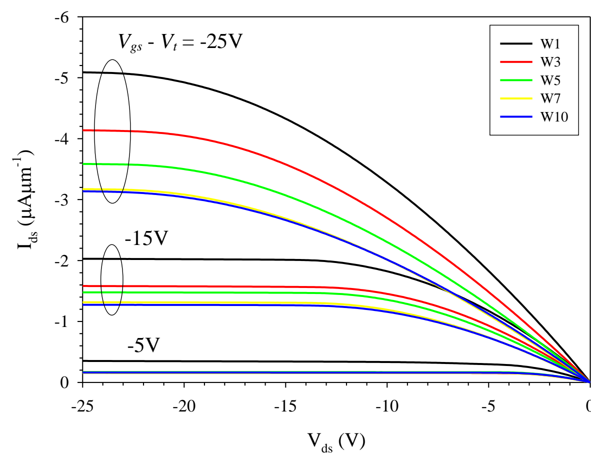


Figure 5.23. I_{ds}-V_{ds} characteristics for devices from various wafers.

5.4.7 Self-Heating

Owing to the considerable length of the majority of devices in the batch, saturation drain currents were typically quite low and only began to exceed 100μAμm^{-1} as the channel length dropped below 10μm. Devices on wafer 1 with a channel length of 300μm typically dissipated just 0.15mWμm^{-1} under large gate and drain bias. This meant that self-heating was not usually an issue that needed to be considered during

electrical characterisation of the long devices. However, the highest germanium content wafers did show appreciable self-heating when large biases were applied to the shorter devices. This is evident in figure 5.24, as the drain current enhancement enjoyed by the strained silicon devices deteriorates with reducing channel length, until the silicon control devices have a higher saturation current. It should be noted that the apparent degradation of the strained silicon devices at short channel lengths may also have affected this result. Data for the 3μm devices is not included because they were unable to withstand a drain bias of -30V without breakdown occurring.

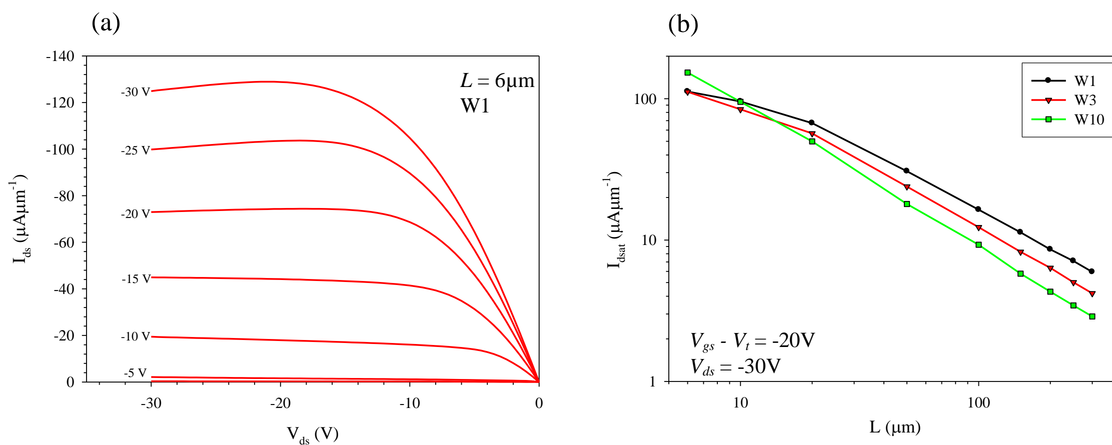


Figure 5.24. (a) pMOSFET I_{ds}-V_{ds} characteristic displaying the effects of self-heating. (b) Variation of saturation drain current with channel length.

Unfortunately, the pulsed measurement system described in chapter 4 was not capable of applying biases greater than 10V in magnitude without extensive modification. At $V_{gs} = V_{ds} = -10V$, self-heating was negligible for all devices, so pulsed measurements were of no use for this batch.

5.4.8 Velocity Saturation

Whilst the devices in batch k2295 were clearly not short enough to exhibit velocity overshoot effects, it was expected that velocity saturation should begin to limit the

behaviour of the shorter devices at high longitudinal electric fields. Evidence of this had already been seen in the saturation transconductance measurements (section 5.4.5). To compare the velocity saturation characteristics of the strained silicon devices to the silicon controls, a similar method to that used by Andrieu *et al.* [15] was employed. They plot the ratio of saturation drain current for pseudomorphic SiGe channel devices compared to silicon controls against the ratio of maximum transconductance at low V_{ds} for different channel lengths, expecting to see the ratio of saturation current decrease with channel length but the transconductance ratio (which is directly related to carrier mobility), to remain the same. To avoid complications with the peak transconductance possibly occurring during population of the SiGe, here the ratio of mobilities at a given gate overdrive was plotted directly instead, where mobility has been calculated from the simple expression for drain current (equation 5.2) and corrected for source/drain resistance and effective length. The results of this investigation are shown in figure 5.25.

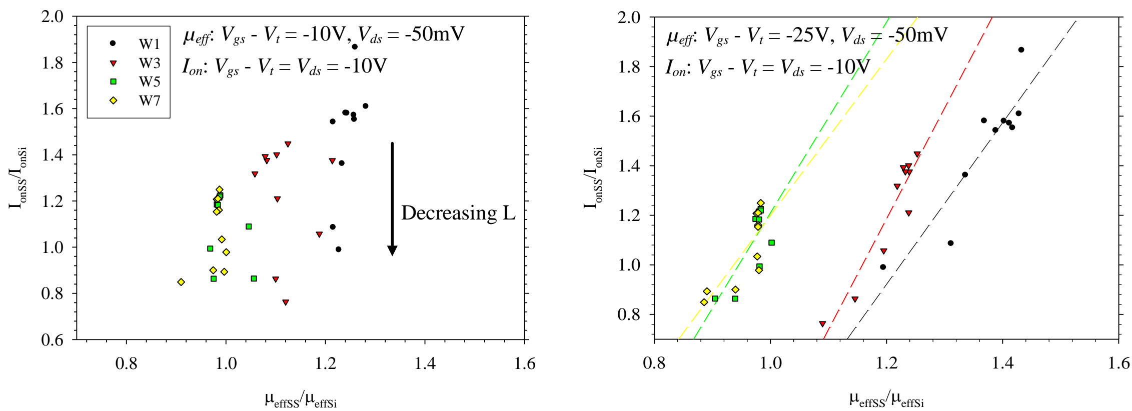


Figure 5.25. Ratio of saturation drain current against ratio of mobility at two different vertical effective fields.

The ratio of the mobilities was calculated under two different bias conditions: firstly for V_{gs} - V_t = -10V, where considerable population of the SiGe would have been occurring for the higher germanium content wafers and secondly for V_{gs} - V_t = -25V, where the majority of carriers would be confined to the strained silicon cap. In both cases, the drain current was extracted under the relatively modest bias of V_{gs} - V_t = V_{ds} = -10V. The reason for this was to reduce the effect of self-heating but also because the shortest devices with $L_{eff} \approx$ 2µm were unable to sustain a drain bias much greater than this before breaking down. The results show that the ratio of saturation drain currents remains approximately the same for a cluster of device lengths (20 - 300µm) but decreases for written device lengths of 3, 6 and 10µm, implying an earlier onset of velocity saturation for the strained silicon devices. This appears to be independent of the degree of strain.

For V_{gs} - V_t = -10V, the carrier mobility ratio displays no clear dependence on gate length as expected, however when the gate overdrive is increased to -25V it becomes evident that the ratio of the strained silicon to silicon control mobilities is decreasing with gate length as L drops to 20µm and below. For clarity, the mobility enhancement of wafer 1 over the control is re-plotted as a function of gate length (figure 5.26). This appears to confirm the theory that excessive grouping of boron has degraded the mobility of the strained silicon layer for short gate lengths. However, if this were the case, then it would be expected that the control would also have a reduced mobility at shorter channel lengths. Figure 5.27 confirms that carrier mobility is apparently reduced in both the strained silicon and the silicon devices, but the higher mobility strained silicon channel suffers a heavier loss of mobility, accounting for the reduced mobility ratio.

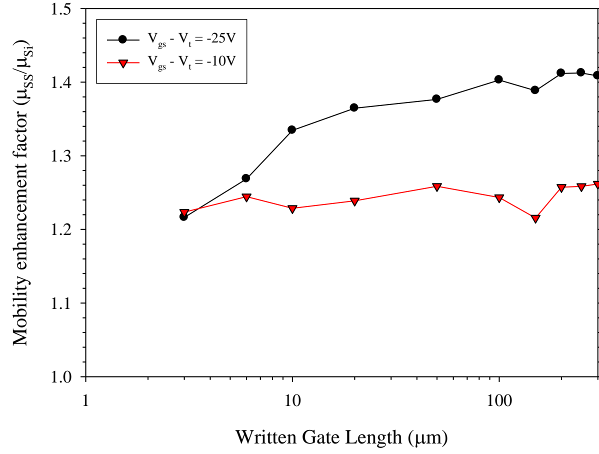


Figure 5.26. The ratio of mobilities between devices on wafers 1 and 10 as a function of gate length.

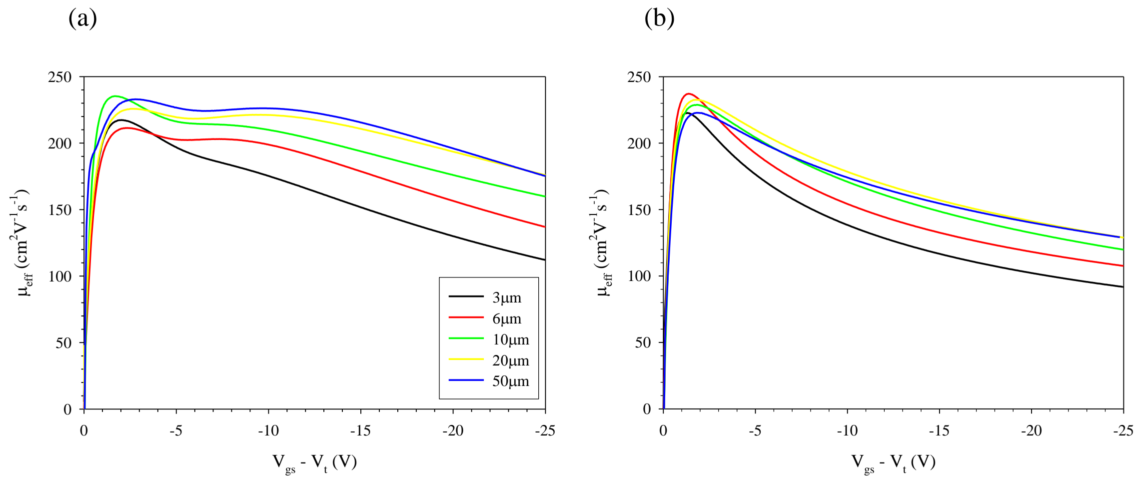


Figure 5.27. Mobility as a function of gate overdrive at different gate lengths for (a) wafer 1 and (b) wafer 10.

5.4.9 Impact Ionisation

Using n^- doped strained silicon resistors, Waldron *et al.* [(16)] have found that the impact ionisation multiplication coefficient, M-1 = I_b/I_s, for electrons in strained silicon (on $Si_{0.8}Ge_{0.2}$) is almost two orders of magnitude higher than that of silicon. Here I_b is the body or substrate current and I_s is the source current. In order to investigate impact

ionisation for holes in strained silicon, the body and source currents of L = 10μm pMOSFETs from wafers 5, 7 and 10 were measured in deep saturation. However, in order to ensure that the majority of holes would be contained within the strained silicon channel, a large vertical effective field was required. This reduces the impact ionisation because it reduces the carrier velocity for a given longitudinal field, due to mobility degradation. A trade-off was therefore required and it was decided to investigate the devices under a vertical field of approximately 0.2MVcm^{-1}, which meant that for wafers 5 and 7, the strained silicon channel would be strongly populated relative to the SiGe buffer.

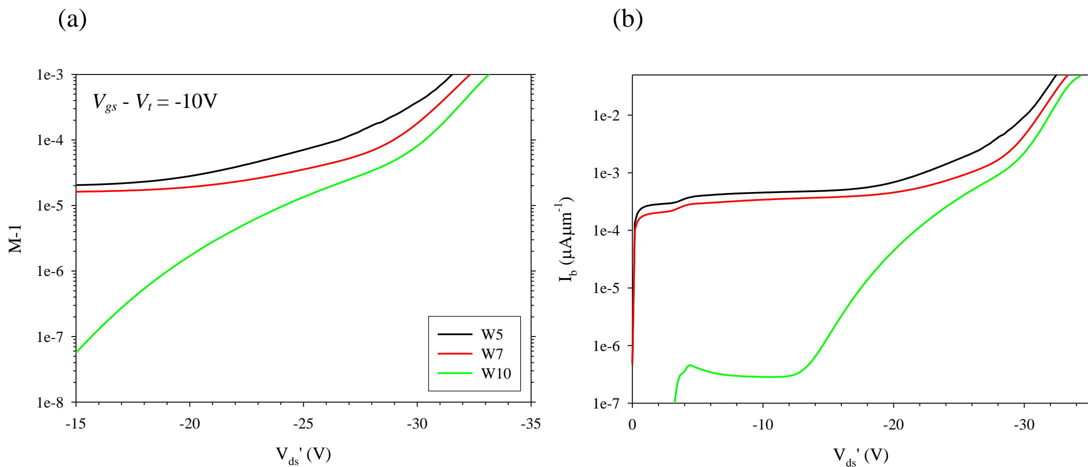


Figure 5.28. (a) Impact ionisation multiplication coefficient and (b) substrate leakage current for wafers from k2295.

From figure 5.28(a) it may be seen that the impact ionisation multiplication coefficient appears to increase with increasing strain, being approximately one order of magnitude higher for the device on wafer 5 than for the silicon control. Some care must taken in drawing conclusions about impact ionisation in the strained silicon devices because the substrate current is several orders of magnitude higher than for silicon, even under small longitudinal fields when impact ionisation is not a factor,

due to the existence of threading dislocations (section 5.4.5). In addition to this, the issue of self-heating must be considered. As the drain bias increased beyond -30V, the dissipated power in the devices approached 1mWμm^{-1}, which is beyond a level at which self-heating can be considered negligible. Impact ionisation in silicon is only weakly dependent on temperature, decreasing slightly as temperature increases. However, for their n-type strained silicon resistors, Waldron *et al.* find a significant increase in impact ionisation with increasing temperature. This they attribute to the reduction of the strained silicon energy gap with increasing temperature, which overcomes the effect of the reduction in mean free path. Owing to the complications of excessive body current and self-heating in these devices, it is difficult to reach any definite conclusions about impact ionisation for holes in strained silicon for this batch.

5.5 k2334 Specification

The strain-relaxed virtual substrates in batch k2334 had a 3μm linearly graded region and a further 1μm of SiGe at constant composition grown on (100) silicon substrates. Subsequent to LEPECVD growth, a further 0.5μm of constant composition SiGe at 10^{17}cm^{-3} n-type was grown by SS-MBE using antimony as the dopant. 50nm spacer layers were grown between the doping and the strained silicon layers, so that the doping was approximately 10^{16}cm^{-3} at the semiconductor surface. The wafer investigated here had a terminating composition of 20% germanium, with a strained silicon cap thickness of 17nm. The silicon control also had an epitaxially grown doped region.

Device fabrication proceeded at Southampton University. Creation of the gate oxide was at 800ºC and quasi-static C-V measurements on MOS capacitors revealed that the

oxide thicknesses was 4.5 ± 1nm for the $Si/Si_{0.2}Ge_{0.8}$ wafer and 5.7 ± 1nm for the silicon control, reflecting the slower oxidation rate of strained silicon. The source and drain implants included creation of low doped drain structures and were activated with a 1020ºC RTA. The gate was n^+ polysilicon. A full titanium salicide process was used and the shortest MOSFETs had written gate lengths of 100nm, although fabrication was optimised for the L = 0.25μm pMOSFETs. Further details of the fabrication procedure are given by Temple *et al.*.(17)

5.6 k2334 Electrical Characterisation

5.6.1 Mobility

Much of the basic electrical characterisation of devices in this batch, including the effective mobility dependence on vertical field, has been performed by Temple *et al.* as part of a different investigation.(17) However, due to its relevance to the further investigations presented here, the mobility extraction was repeated. Again this was accomplished by a combination of I-V, quasi-static C-V and split C-V measurements on the largest devices provided, which in this case were 10 × 100μm pMOSFETs. It was also found that R_{sd} = 1107 ± 200Ωμm, 1808 ± 200Ωμm and that ΔL = -6 ± 10nm, 4 ± 10nm for the strained and unstrained wafers respectively. Figure 5.29 shows that the strained silicon device displayed a substantial enhancement over the universal mobility curve for the entire range of vertical fields probed here. At 0.8MVcm^{-1} the effective hole mobility is approximately 40% higher than the universal mobility found by Takagi *et al.*. There is no evidence of parasitic conduction, as expected for strained silicon devices with a thin gate oxide and relatively low germanium content in the underlying buffer. The silicon control wafer for this batch was found to have a very poor mobility at moderate fields, in agreement with the work of Temple *et al.*. It

appears that the doping in the control was higher than that in the strained silicon wafer. The reason for this is unclear, but is confirmed by the threshold voltages of the devices, which were typically ~ 1V larger for the silicon pMOSFETs.

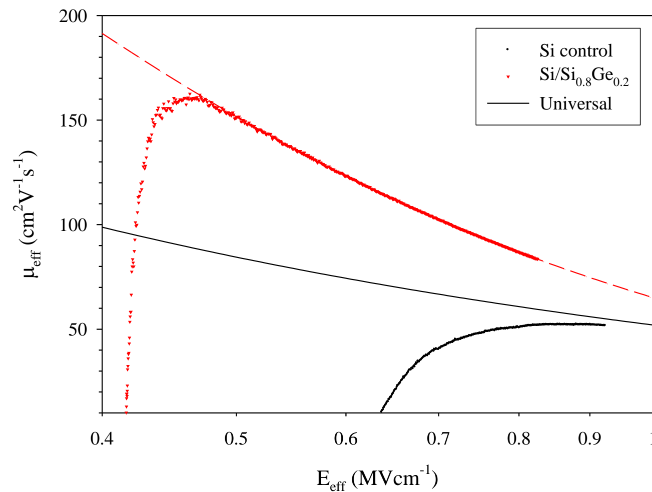


Figure 5.29. Effective mobility as a function of effective vertical field for 10 × 100µm pMOSFETs. The dashed line represents an empirical fit to the strained silicon mobility.

A curve was fitted to the experimentally determined mobility for the strained silicon device, again using equation 5.1. This yielded a = 0.249MVcm^{-1}, in close agreement with the silicon control of batch k2295. The extracted low field mobility was 579cm^{2}V^{-1}s^{-1} and b was determined as 1.49, indicating a stronger dependence on electric field than was evident in any of the wafers from k2295. The low field mobility is predicted to be approximately double that of the universal mobility for holes in silicon, which somewhat exceeds the prediction of Oberhüber *et al.* for strained silicon on a 20% buffer, although it should be noted that no work has been carried out to confirm the composition of the SiGe buffer. Despite the large reduction in mobility with increasing field, the strained silicon pMOSFET is predicted to retain

an enhancement of approximately 25% over the universal mobility at a vertical effective field of 1.0MVcm^{-1}.

5.6.2 Carrier Velocity

In order to investigate how close to the thermal limit short channel strained silicon pMOSFETs are operating, carrier velocity was extracted in two different ways as described in section 3.5.11. The initial results of this investigation are displayed in figure 5.30.

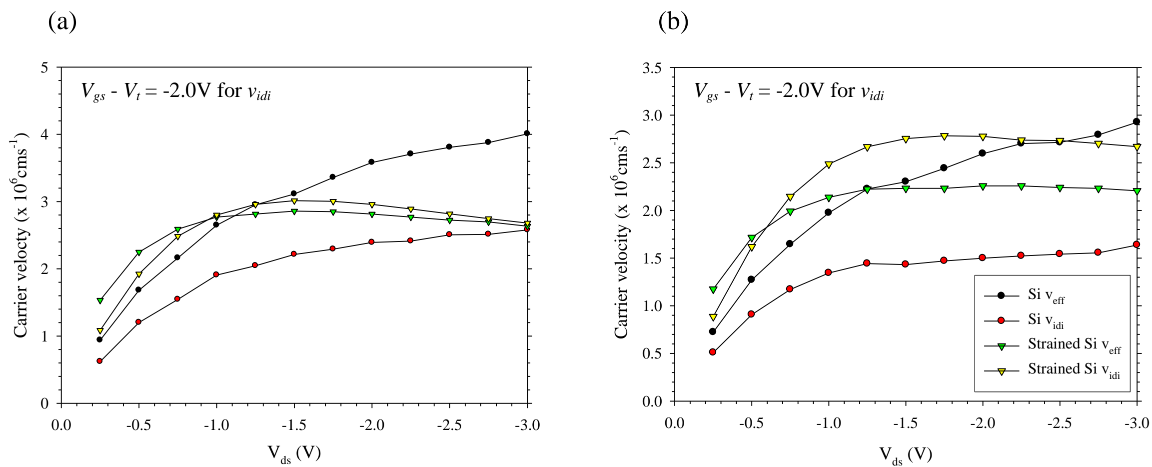


Figure 5.30. Hole velocity as extracted by two different methods for (a) L = 100nm and (b) L = 150nm strained and unstrained silicon pMOSFETs.

It may be seen that for the silicon control devices, the average channel velocity extracted using the familiar $v_{eff} = g_{mi}/WC_{ox}$ relation is approximately twice that of the velocity near the source, extracted using the method proposed by Lochtefeld *et al.*.[18] In the case of the 100nm silicon pMOSFET, both the average channel velocity and the velocity near the source are lower than the saturation velocity even under high drive conditions, indicating that at these gate lengths there is still an advantage to be

realised by improving the carrier mobility. It should be noted that the carrier velocities v_{idi} and v_{eff} are not directly comparable (except at $V_{ds} \approx V_{gs} - V_t$), because v_{eff} is extracted from the peak transconductance, which occurs for different values of $V_{gs} - V_t$ as V_{ds} is varied.

The results from the strained silicon devices at first appear highly unusual, since at moderate longitudinal fields the carrier velocity drops below that of the silicon control and the average channel velocity actually appears to be less than the injection velocity. This is the result of substantial self-heating present in these devices, which will affect the transconductance more than the drain current and hence lead to a greatly reduced apparent carrier velocity. The reason that the apparent average channel velocity lies closer to the injection velocity for the 100nm device than the 150nm, when the inverse might be expected due to the presence of increased self-heating, is likely to be DIBL. The 100nm strained silicon devices had a much higher DIBL (420mV/V) than their 150nm counterparts (100mV/V). The standard extraction of carrier velocity from the peak transconductance is sensitive to DIBL and can lead to high carrier velocities when DIBL is poor. For this reason the extracted average carrier velocities in particular must always be treated with caution, even for the silicon devices, where the DIBL was somewhat less. In order to counter the effect of self-heating, pulsed measurements were performed on the strained silicon devices using the equipment described in chapter 4.

Figure 5.31 shows that a surprisingly large degree of self-heating was observed, given that the terminating composition of the virtual substrate was very similar to that of the devices previously measured (see section 4.5) and that the maximum power dissipated

was 0.6mWμm^{-1}. This is therefore likely to be due to the increased thickness of both the graded layer and the constant composition buffer. Using equation 4.1 to approximately calculate the thermal resistance of the devices reveals $R \approx 21.7$KmW^{-1} for the 10μm wide devices, which is substantially higher than that that predicted for the ST Microelectronics devices (section 4.4).

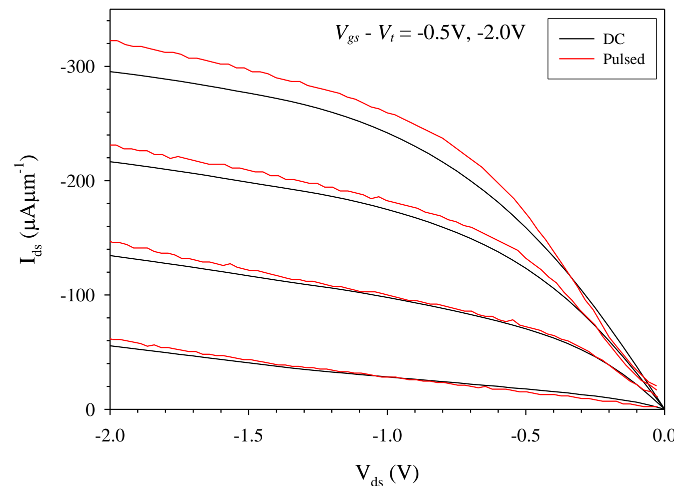


Figure 5.31. Comparison of pulsed and quasi-static measurements of a 100nm strained silicon pMOSFET from batch k2334.

From the pulsed measurements, the corrected carrier velocities near the source were relatively easily calculated. A problem became apparent, however, when attempting to extract transconductance characteristics in the absence of self-heating. Because the drain current as measured by the pulsed system could only take discrete values, relatively widely spaced, a large error was introduced into the extracted transconductance. In an attempt to overcome this problem, the I_{ds}-V_{gs} data were smoothed using *SigmaPlot* version 8.0. As may be seen from figure 5.32, this process was only moderately successful and consequently an error of approximately ± 10% in the extracted peak transconductances must be assumed.

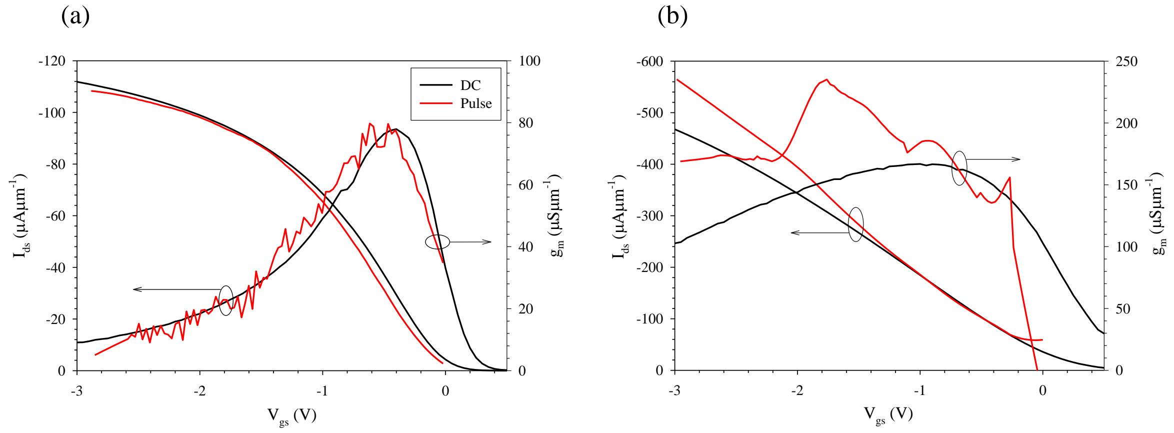


Figure 5.32. Comparison of pulsed and quasi-static transconductance and I_{ds}-V_{gs} curves for (a) V_{ds} = -0.25V and (b) V_{ds} = -2V.

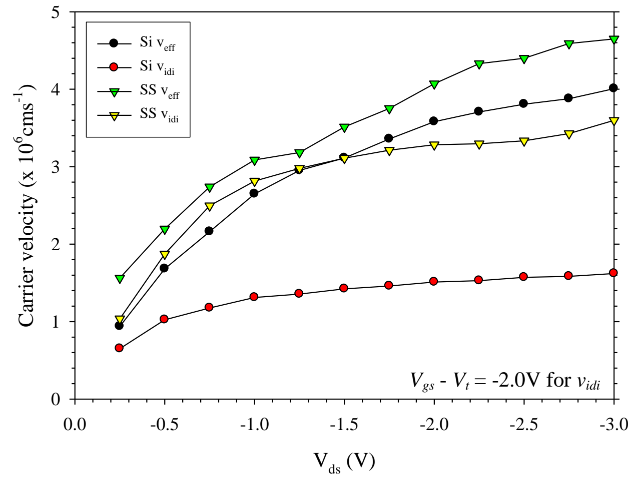


Figure 5.33. Hole velocity as extracted by the two different methods, using the data from pulsed measurements for the 100nm strained silicon device.

Although the errors in the hole velocity data for the strained silicon device shown in figure 5.33 were thought to be large, particularly for the standard v_{eff} extraction, the results show that the injection velocity of holes into the channel is well below the thermal limit (~ 1.4 × 10^7 cms^{-1}) at a gate length of 100nm. This indicates that strained silicon is likely to be able to provide a performance enhancement over standard silicon pMOS for several technology generations to come. The hole velocity at the source, v_{idi}, is however much higher in the strained silicon device than the

control. In reality this comparison is for carriers that have already been accelerated over a short section of the channel, so it is likely that the elevated mobility of the strained silicon device has exaggerated the difference in injection velocity. The much greater DIBL in the strained silicon device is also likely to account for some of the difference.

5.6.3 Impact Ionisation

Further to the work presented in section 5.4.9, impact ionisation in devices of batch k2334 was investigated. The high material quality of these wafers meant that the problem of body current due to factors other than impact ionisation was largely overcome. Figure 5.34 shows that the strained silicon devices exhibited over six decades of I_{ds} variation, whilst the silicon control had seven decades, although a positive V_{gs} was required to fully turn off the strained silicon devices. In addition, thanks to the thin gate oxides and n-type doping of the SiGe buffer, parasitic conduction was not an issue for batch k2334 and thus the impact ionisation of holes in strained silicon could be studied for a small gate overdrive, removing the complication of self-heating. 10 × 10μm pMOSFETs were investigated under a gate overdrive of -0.5V.

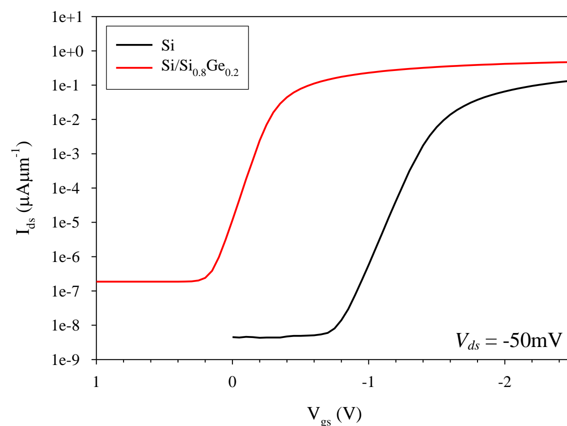


Figure 5.34. Comparison of turn-on characteristics of 10 × 10μm pMOSFETs.

(a) (b)

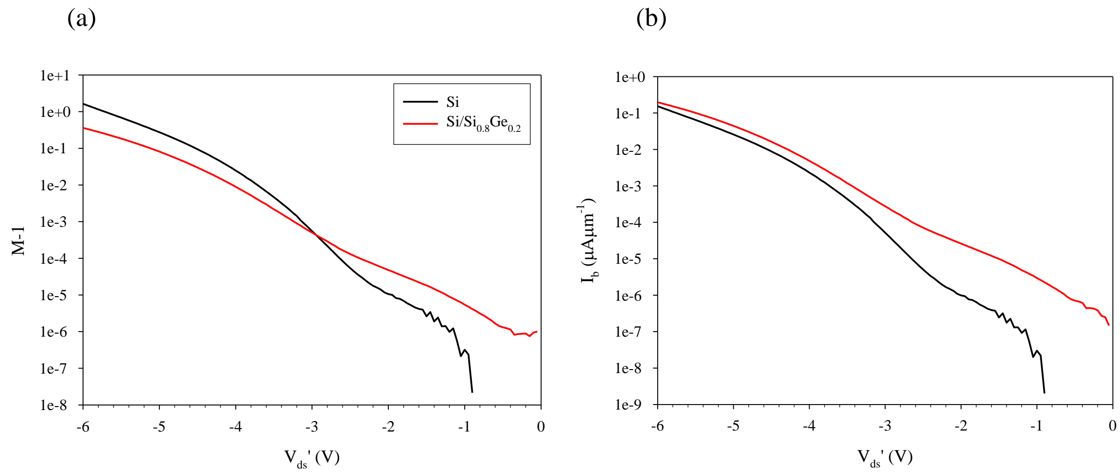


Figure 5.35. (a) Impact ionisation multiplication coefficient and (b) the body current as a function of intrinsic source drain voltage.

Figure 5.35(a) appears to suggest that the strained silicon device suffers from an earlier onset of impact ionisation. However, it is again unclear that the increasing body current is due to impact ionisation and not simply a result of elevated junction leakage at increased drain bias. Figure 5.35(b) suggests that impact ionisation in the control begins when $V_{ds}' \approx$ -2.5V, as there is a sharp increase in the body current. The body current in the strained silicon device is always higher than that in the control but this appears to be because of junction leakage, due the smaller bandgap of SiGe. There is a marked increase in body current at $V_{ds}' \approx$ -3V, implying that in fact the onset of impact ionisation is delayed in the strained silicon device. If we consider that impact ionisation does not feature strongly until such a drain bias is reached, figure 5.35(a) indicates that impact ionisation for holes in strained silicon is less than in bulk silicon. By simply plotting the ratio of the drain current to the source current (figure 3.36), it is apparent that impact ionisation for holes may be substantially reduced in a strained silicon device. This is a surprising result in light of the reduced energy gap and increased mean free path of strained silicon, both of which would tend to increase

impact ionisation. Increased doping in the silicon control may have contributed to junction leakage, and if the doping was also higher at the surface then the resultant increase in lateral electric field would have resulted in more impact ionisation. However, the uppermost ~ 70nm of both the control and the strained silicon wafer consisted of undoped material. The possibility therefore exists that the mean free path in strained silicon is sufficiently high that a substantial proportion of the hot carriers near the drain end of the channel are collected before colliding with the lattice. Further investigation is needed.

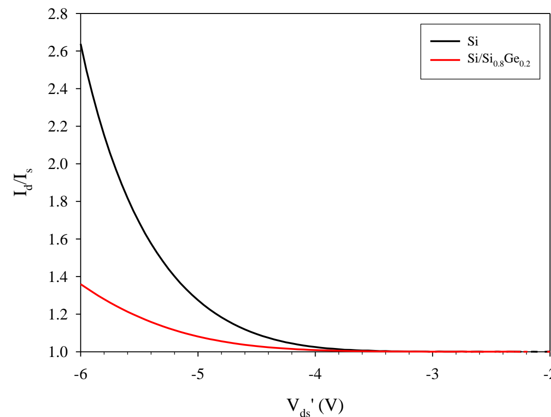


Figure 5.36. Ratio of drain current to source current for the 10 × 10μm pMOSFETs.

5.7 ECOPRO Devices

Under the ECOPRO program, a batch of twenty strained silicon wafers were to undergo a 0.5μm device fabrication at NMRC in Cork, Ireland. The batch was split into eight wafers for nMOSFETs and twelve wafers for pMOSFETs. Due to NMRC's inexperience of device fabrication on virtual substrates, the terminating compositions were restricted to 20 and 30% germanium. The wafers for the nMOSFETs had a boron doped SiGe buffer grown on top of the virtual substrates by SS-MBE, followed by the strained silicon cap. Due to the difficulties of n-type doping during MBE growth,[18] the wafers for the pMOSFETs were doped with antimony via ion

implantation. Half of the wafers were implanted prior to SS-MBE growth of the strained silicon layer, with the other half implanted subsequent to growth.

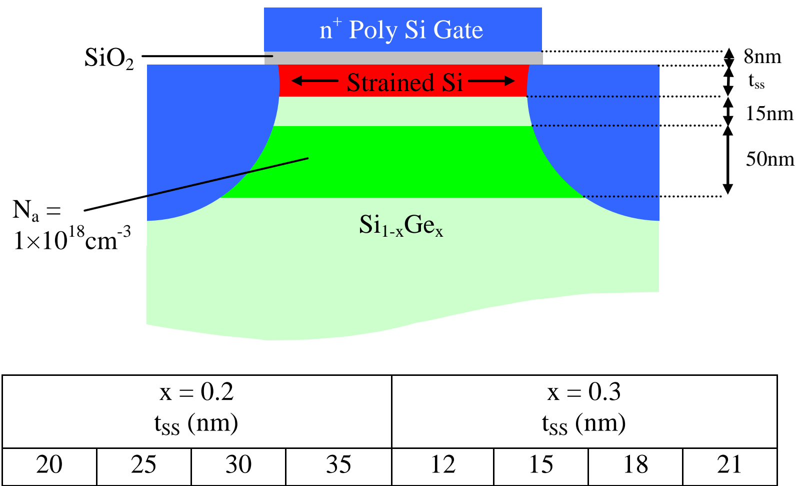
5.7.1 Design of Devices

Device simulations were used to decide the doping concentrations required to produce suitable threshold voltages. Simulation is discussed in chapter 6, but in this chapter some predicted characteristics are included for comparison with the results from measurements of the devices. Information that was given to aid the device simulations was as follows:

- The thinnest gate oxide layer that could be created by thermal oxidation was 8nm.
- The interface trap density at the Si/SiO_2 interface was likely to be in the region of $2\times10^{11}cm^{-2}eV^{-1}$ for the silicon control wafers.
- A typical antimony doping profile resulting from ion implantation would have a peak concentration at a depth of 32nm and a straggle of 42nm, although these figures were given for ion implantation into silicon.

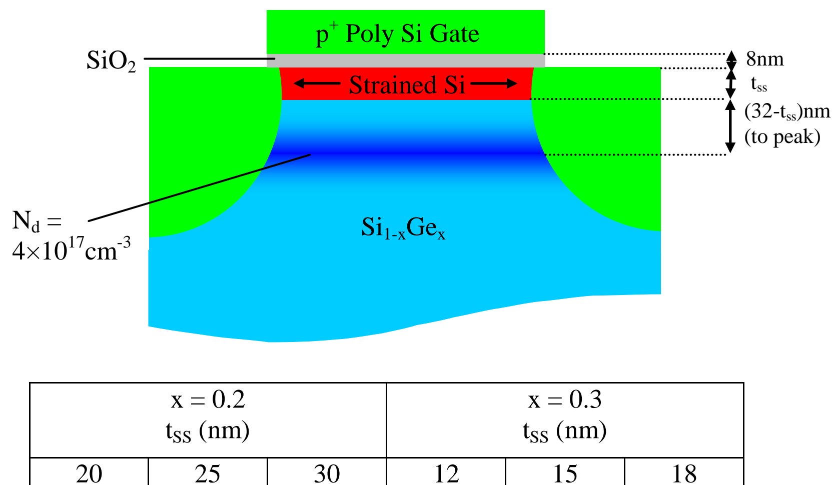
The thinnest possible gate oxides were requested in order to reduce parasitic conduction as far as possible in the pMOSFETs. In addition, the strained silicon layers were to be made as thick as possible. The thinnest strained silicon layers were therefore grown at the equilibrium critical thickness predicted by Samavedam *et al.*,(19) whilst the thickest layers were 175% of this value. Simulations were performed with the aim of producing devices with a threshold voltage close to ± 0.5V, in line with the roadmap value for MOSFETs of this geometry. It was predicted that in the absence of any doping beyond the background of the growth chambers, the threshold voltages would lie close to zero. Additional doping was therefore necessary

but it was important to remove it from the channel as much as possible in order to avoid carrier scattering at ionised impurities. Simulations showed that a doping slab could effect strong control of the channel provided it was within ~ 50nm. Previous experience suggested that a separation of 15nm was necessary to protect the carrier mobility.[20]



x = 0.2 tss (nm)				x = 0.3 tss (nm)			
20	25	30	35	12	15	18	21

Figure 5.37. Specification for the Si/SiGe nMOSFETs with the as-grown strained silicon thickness shown.



x = 0.2 tss (nm)			x = 0.3 tss (nm)		
20	25	30	12	15	18

Figure 5.38. Specification for the Si/SiGe pMOSFETs with the as-grown strained silicon thickness shown.

The finalised designs for the two types of MOSFET are shown in figures 5.37 and 5.38. It should be noted that at the time of performing the simulations, there was great uncertainty concerning the band structure and the band discontinuities used were smaller than the work presented in chapter 6 would suggest.

5.7.2 I-V Measurements

Due to serious delays in device fabrication at NMRC, only three half wafers were available for characterisation. These wafers were 20% germanium virtual substrates with as-grown strained silicon thicknesses of 20nm (wafer 7), 25nm (wafer 8) and 35nm (wafer 10) and had undergone a 0.5μm nMOSFET fabrication process. No silicon control was available for comparison.

Unfortunately, two of the three half wafers supplied were largely electrical inactive. The following results are therefore restricted to measurements of the remaining half wafer, that with the 20nm strained silicon layer. It was apparent that on this wafer, none of the devices with a written gate length below 1μm displayed acceptable transistor action. As may be seen from figure 5.39, it was not possible to turn-off the L = 0.75μm nMOSFETs, as the I_{on}/I_{off} ratio was less than one decade, with the L = 0.5μm nMOSFETs being even worse. It is likely that punchthrough was occurring for these relatively short devices. For $L \geq 1$μm, the typical I_{on}/I_{off} ratio was approximately 5 decades. The extracted threshold voltage of the 0.75μm device is 0.76V, which is somewhat larger than the 0.55V predicted by device simulations. The threshold voltages of the longer devices were ~ 0.85V, but the simulations had been performed for short devices.

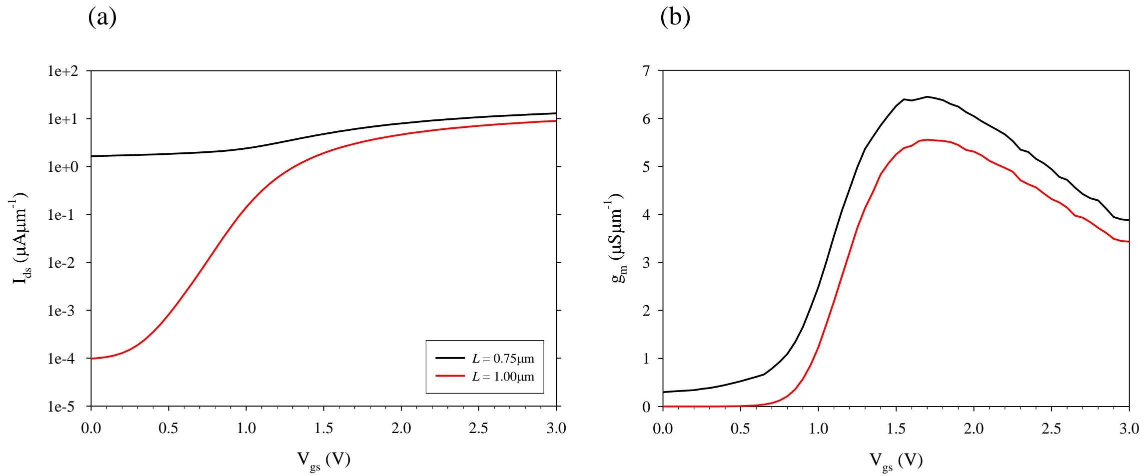


Figure 5.39. (a) Turn-on characteristics and (b) transconductance for nMOSFETs from the ECOPRO batch.

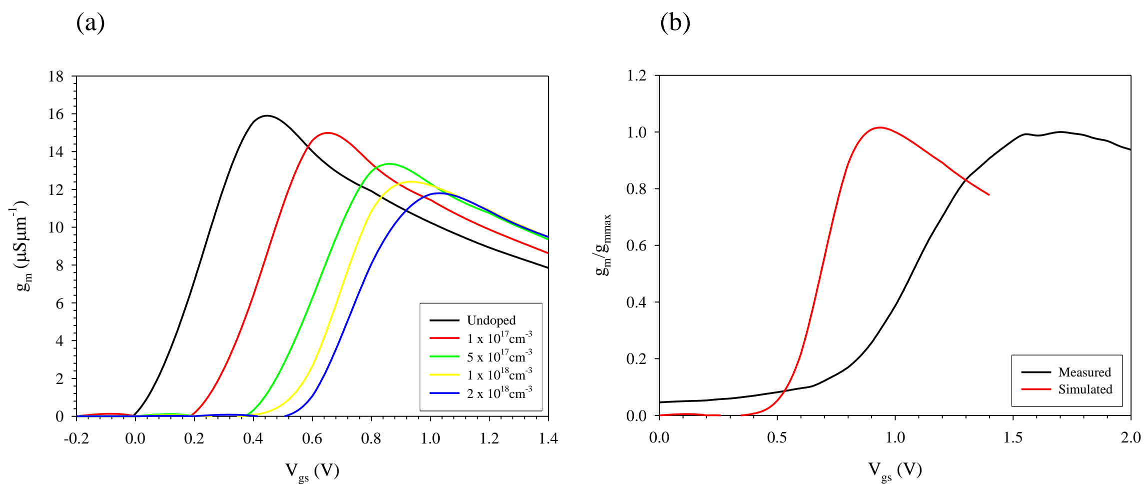


Figure 5.40. (a) Simulated transconductance for L = 0.75µm nMOSFETs with different doping concentrations in the buffer. (b) Comparison of simulated and measured transconductance for L = 0.75µm nMOSFETs.

5.7.3 C-V Measurements and Mobility

High and low frequency C-V measurements revealed a very high interface trap density (> 10^{12}cm^{-2} at midgap). The interface traps are clearly visible in the quasi-static C-V measurement (figure 5.41) as a peak before the onset of inversion. The

apparent doping concentration was approximately $10^{18}cm^{-3}$, although due to the false rise in doping concentration at the surface of the device which is a feature of the C-V technique, it was not possible to extract a dopant profile.

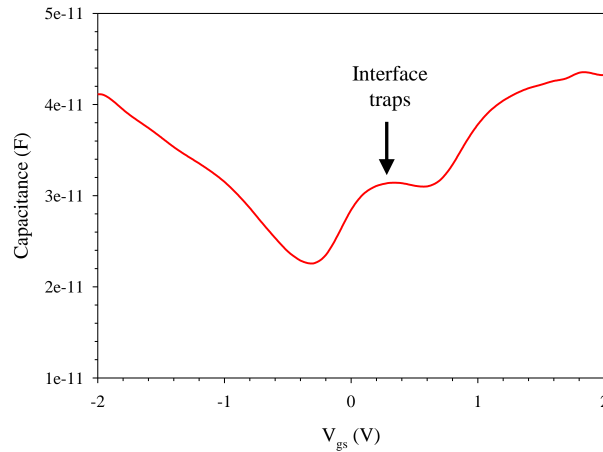


Figure 5.41. Quasi-static C-V measurement of a 70 × 100μm capacitor, showing a high interface trap density.

The oxide thickness extracted from the quasi-static C-V measurements was approximately 6nm, which was somewhat thinner than expected. I-V and C-V measurements were combined to extract the effective mobility as a function of effective field for a 70 × 100μm nMOSFET (figure 5.42). It should be noted that no correction was made for R_{sd} or L_{eff}, although at this channel length, the effect of these parameters should be negligible. Since no control was available, the universal electron mobility (9) and the results of Rim *et al.* (21) for strained silicon on $Si_{0.8}Ge_{0.2}$ are included in figure 5.42 for comparison. It is apparent that at high vertical fields, the electron mobility began to exceed the universal curve; however it was not possible to probe higher fields before breakdown of the gate oxide occurred. The poor mobility at moderate fields may be attributable to scattering at ionised impurities, as it seems likely that the doping concentration in the channel region was higher than

expected. The conduction band offsets used in the device simulations prior to device fabrication were smaller than in reality, which should have resulted in the simulated threshold voltage being larger than that measured in reality. The fact that the measured V_t was larger suggests an elevated p-type doping concentration. This would be surprising in light of the apparent punchthrough problems of the shortest devices, although this may be explained by problems with the source/drain implantation. The characteristics of the ECOPRO wafers are certainly disappointing, given that SS-MBE growth on LEPECVD virtual substrates provided record hole mobilities in batch k2334, with the only difference being the choice of fabrication facility.

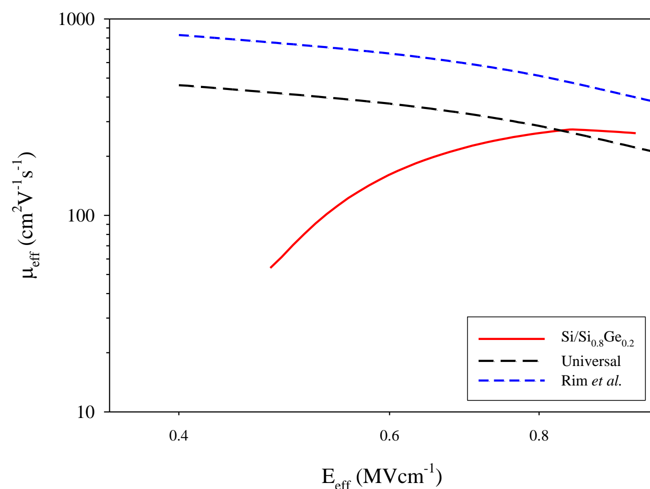


Figure 5.42. Extracted electron mobility as a function of effective field for a 70 × 100µm nMOSFET.

References

1. H. von Känel *et al.*, "Fast Deposition Process for Graded SiGe Buffer Layers", *Japan J. Applied Physics*, vol. 39, no. 4B, pp. 2050-2053, 2000.
2. "Plasma Processing of Materials: Scientific Opportunities and Technological Challenges", National Academy Press, pp. 15, 1991.
3. S. H. Olsen *et al.*, "Strained Si/SiGe n-Channel MOSFETs: Impact of Cross-Hatching on Device Performance", *Semicond. Sci. Technol.*, vol. 17, no. 7, pp. 655-661, 2002.

4. M. T. Lin, R. J. Jaccodine and T. J. Delph, “Planar Oxidation of Strained Silicon Substrates”, *Journal of Materials Research*, vol. 16, no. 3, pp. 728-733, 2001.
5. M. M. Rieger and P. Vogl, “Electronic-Band Parameters in Strained $Si_{1-x}Ge_x$ Alloys on $Si_{1-y}Ge_y$ Substrates”, *Phys. Rev. B*, vol. 48, no. 19, pp.14276-14287, 1993.
6. S. S. Iyer *et al.*, “Growth Temperature Dependence of Interfacial Abruptness in Si/Ge Heteroepitaxy Studied by Raman Spectroscopy and Medium Energy Ion Scattering”, *Appl. Phys. Lett.*, vol. 54, no. 3, pp. 219-221, 1989.
7. C. G. Ahn *et al.*, “Oxidation-Induced Traps near SiO_2/SiGe Interface”, *J. Appl. Phys.*, vol. 86, no. 3, pp. 1524-1547, 1999.
8. R. Oberhüber, G. Zandler and P. Vogl, “Subband Structure and Mobility of Two-Dimensional Holes in Strained Si/SiGe MOSFET’s”, *Phys. Rev. B*, vol. 58, no. 15, pp. 9941-9947, 1998.
9. S. Takagi, A. Toriumi, M. Iwase and H. Tango, “On the Universality of Inversion Layer Mobility in Si MOSFET’s: Part I-Effects of Substrate Impurity Concentration”, *IEEE Trans. Elec. Dev.*, vol. 41, no. 12, pp. 2357-2362, 1994.
10. J. T. Watt and J. D. Plummer, “Universal Mobility-Field Curves for Electrons and Holes in MOS Inversion Layer”, *Proc. VLSI Symp.*, pp. 81, 1987.
11. T. Mizuno *et al.*, “Physical Mechanism for High Hole Mobility in (110)-Surface Strained- and Unstrained-MOSFETs”, *IEEE IEDM Tech. Dig.*, 2003.
12. T. Manku and A. Nathan, “Lattice Mobility of Holes in Strained and Unstrained $Si_{1-x}Ge_x$ Alloys”, *IEEE Elec. Dev. Lett.*, vol. 12, no. 12, pp. 704-706, 1991.
13. S. Kaya *et al.*, “Indication of Velocity Overshoot in Strained $Si_{0.8}Ge_{0.2}$ p-Channel MOSFETs”, *Semicond. Sci. Technol.*, vol. 15, no. 6, pp. 573-578, 2000.
14. S. Roy, *SiGe HMOS Consortium MOSFET Training Course*, University of Glasgow, 2002.
15. F. Andrieu *et al.*, “SiGe Channel p-MOSFETs Scaling-Down”, *33rd European Solid-State Device Research Conference Proceedings*, pp. 267-270, 2003.
16. N. S. Waldron *et al.*, “Impact Ionization in Strained-Si/SiGe Heterostructures”, *IEEE IEDM Conference Proceedings*, 2003.
17. M. P. Temple *et al.*, “Enhanced p-MOSFET Performance using Strained-Si on SiGe Virtual Substrates Grown by Low Energy Plasma Enhanced Chemical Vapour Deposition”, *unpublished*.
18. A. Lochtefeld, I. J. Djomehri, G. Samudra, D. A. Antoniadis, “New Insights into Carrier Transport in n-MOSFETs”, *IBM J. Res. & Dev.*, vol. 46, no. 2/3, pp.347-357, 2002.
19. S. B. Samavedam *et al.*, “Relaxation of Strained Si Layers Grown on SiGe Buffers”, *J. Vac. Sci. Technol. B*, vol. 17, no. 4, pp. 1424-1428, 1999.
20. T. J. Grasby, *private communication*.
21. K. Rim *et al.*, “Strained Si CMOS (SS CMOS) Technology: Opportunities and Challenges”, *3rd European Workshop on the Ultimate Integration of Silicon*, pp. 73-76, 2002.

Chapter 6

Device Simulations

6.1 Introduction

When designing a new MOSFET, it is important to ensure that the structure is optimised in terms of parameters such as the I_{on}/I_{off} ratio and the peak transconductance. To save time and money on costly iterative fabrication runs, computer simulations are commonly used in research to ensure that the first fabricated device will be close to achieving the specified requirements.

At Warwick, the 2D drift-diffusion device simulator *Medici* [(1)] is used to help in the design of MOSFETs prior to layer growth and fabrication. The full range of device attributes may be adjusted, including doping profiles, layer thicknesses, contact metal types and device dimensions. The primary function of Medici is to solve three partial differential equations self-consistently for the electrostatic potential, ψ, and the electron and hole concentrations, n and p respectively. The three equations are Poisson's equation:

$$\varepsilon \nabla^2 \psi = -q\left(p - n + N_D^+ - N_A^-\right) - \rho_s, \tag{6.1}$$

and two continuity equations:

$$\frac{\partial n}{\partial t} = \frac{1}{q} \nabla \bullet \underline{J}_n - U_n = F_n(\psi, n, p), \tag{6.2}$$

$$\frac{\partial p}{\partial t} = \frac{1}{q} \nabla \bullet \underline{J}_p - U_p = F_p(\psi, n, p). \tag{6.3}$$

Here ρ_s is the surface charge density (due to interface charge and traps), N_D^+ and N_A^- are ionised impurity concentrations and U_n and U_p represent net electron and hole

recombination. The electron and hole current densities, $\underline{J}_n$ and $\underline{J}_p$, are given by the drift-diffusion equations:

$$\underline{J}_n = q\mu_n n\underline{E}_n + kT\mu_n \nabla n, \qquad \underline{J}_p = q\mu_p p\underline{E}_p - kT\mu_p \nabla p. \qquad (6.4)$$

Medici solves the partial differential equations at a number of locations throughout the structure, designated by the user by means of a mesh (see section 6.2). If a series of contact regions (typically source, drain, gate and substrate) are specified, Medici is able to calculate the current through any of them according to the bias applied. In this way, the device parameters may be adjusted until the simulated electrical characteristics are optimised. This procedure was followed in the design of the devices for the ECOPRO program, detailed in section 5.7.

An alternative and less common use for a device simulator is to simulate the electrical characteristics of a device that has already been fabricated in the hope of gleaning some information about unknown physical properties. Since there are many variables that may be adjusted, and different combinations of these variables can produce very similar output characteristics, it is important that as many unknowns as possible are eliminated from the simulation. To this end, the device being simulated is fully characterised so that such quantities as the doping concentration and layer thicknesses are relatively well known.

In this chapter the procedure for setting up a successful simulation is outlined. These principles are then applied to the simulation of long devices from batch k2295, subsequent to the characterisation detailed in chapter 5. Because the band offsets between strained silicon and relaxed SiGe are still not well defined, these simulations

were performed with the aim of providing some clarification. In this work, the heavy parasitic conduction accompanying the exceptionally thick gate oxide layers and high germanium content of the buffers was exploited to extract information on the valence band offset.

6.2 Simulation Procedure

The first stage in creating a successful Medici device simulation is defining the mesh that will be used. Because the equations in section 6.1 are solved for each mesh point, it is important not to make the mesh too fine, as the simulation will require a great deal of processing time to complete. At the same time, it is important to include enough mesh points that the model is an accurate representation of a real world device. The mesh is usually defined in such a way that it is fine in the active region of the device, becoming coarser as the distance from this area increases. Much of the computer's effort is therefore focused on the important device regions, rather than squandering resources on a highly accurate simulation of carrier transport deep in the substrate (for example).

An example mesh is shown in figure 6.1. Once materials and doping profiles have been defined, it is possible to make use of the *regrid* command in Medici, which further refines the mesh according to the user's preference. For example, it is possible to instruct the computer to create additional mesh points wherever the doping gradient exceeds a certain pre-set value, thus helping to avoid discontinuities in the simulation, which can be problematic.

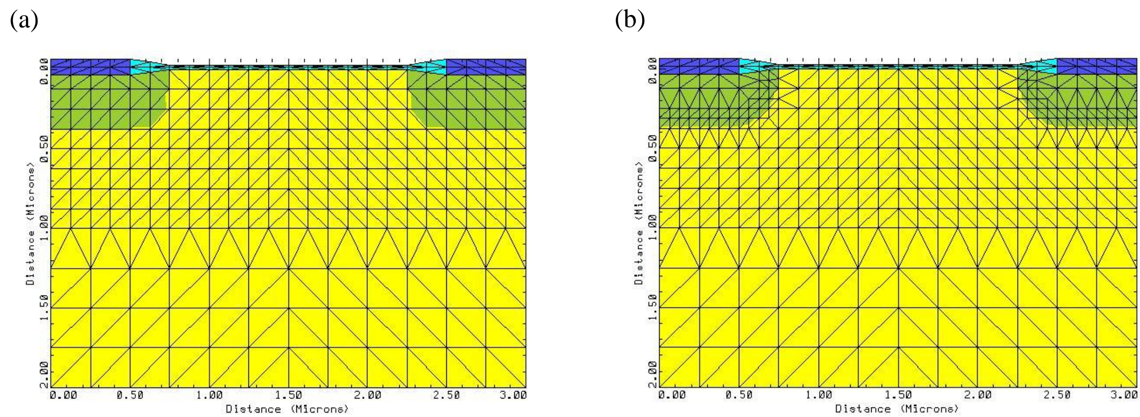


Figure 6.1. Example mesh used for a device simulation (a) before and (b) after refinement around the doping in the source and drain regions.

There are a number of carrier mobility models available in Medici and selection of a suitable one is of great importance. There are several models suitable for simulation of low field mobility, the most simple of which assumes a constant (i.e. field independent) mobility with the most complex including acceptor, donor and carrier-carrier scattering effects. These models are not usually very useful because they do not include the dependence of carrier mobility on electric field strength, either in the vertical or horizontal directions. The familiar universal mobility model is available if vertical field dependence only is required. This is usually sufficient if the simulation involves long channel devices or a small source/drain bias. The most complex mobility models available in Medici also include velocity saturation effects.

There are several ways of fine-tuning a device simulation to improve accuracy. Trap densities may be specified at interfaces between different materials. If the purpose of the simulation is to design a new MOSFET, this will have to be a best guess based on characterised devices that have undergone a similar fabrication process. Series resistances may also be specified. In Medici it is also possible to alter any of the

material properties from their default values. This is useful from the point of view of simulating strained silicon devices. Medici includes data for silicon, germanium and SiGe alloys. In the latter case it is possible to specify whether the energy model for strained or unstrained material should be used. Unfortunately the model for strained SiGe is rather basic and assumes compressive strain. If silicon under tensile strain is required then the user must create a model for the material from scratch.

6.3 Extraction of Si/SiGe Valence Band Offset

6.3.1 Simulation of Silicon Control

Prior to simulation of strained silicon devices, a model of a silicon control pMOSFET from wafer 10 was constructed. This would provide confirmation that the Medici simulation of devices in batch k2295 was accurate, as well as providing a useful baseline to start from when attempting the much more complex simulation of a strained silicon device.

In order to almost entirely eliminate complications arising as a result of short channel effects, long devices of 300μm gate length were simulated. Before commencing the simulation, devices from batch k2295 were thoroughly characterised (sections 5.3 and 5.4) in order that as many device parameters as possible were well known and guesswork would be minimised. For the silicon control device, information on L_{eff} and R_{sd} was employed, together with oxide thickness, interface trap density, doping profile and an estimate of low field mobility. As expected, the source/drain resistance and the effective channel length only had a small effect on the simulation, but as this data was available it was included to improve accuracy as far as possible. Because the oxide thickness was known to vary considerably across the wafer, the measured

device was situated close to the point from which the X-TEM sample had been taken. Quasi-static C-V measurements on the device confirmed the gate oxide thickness. The junction depth of the source drain regions was estimated as 0.3μm from knowledge of the boron implantation energy (50keV).[2]

Medici was used to calculate an I_{ds}-V_{gs} curve for the silicon control model. This was compared to the measured characteristic and one aspect of the simulation was adjusted before computing a new result. For the control, the only parameters that required slight adjustment were the background doping (in order to match the threshold voltage) and the carrier mobility (to match the drain current). During this process it was found that it was useful to compare not only the measured and simulated drain current but also the transconductance, since this highlighted any deviation between the data sets. The simulations did not accurately predict the leakage currents encountered in the devices, but this introduced a negligible error for the purposes of this work.

Medici does not solve Schrödinger's equation and therefore does not include the effects of quantum confinement (section 2.5.7). For a silicon device this is only expected to introduce inaccuracy to the simulation at very large vertical fields. However, for the strained silicon devices it was anticipated that ignoring this effect could lead to incorrect relative occupancy of the strained silicon and SiGe buffer, which would have serious consequences for the extraction of the valence band offset. Medici does include a means of simulating the effects of quantum confinement by artificially widening the energy gap of the semiconductor at the oxide interface. In

the interests of continuity, this effect was included during simulation of the silicon control, although it had a negligible effect on the threshold voltage, as expected.

Since the gate length of the simulated device was 300μm and V_{ds} was -50mV, the longitudinal electric field was negligible. The mobility model used was therefore only required to include vertical field dependence. A problem became apparent when attempting to include the effects of quantum confinement however, because the removal of carriers from the Si/SiO_2 interface resulted in an unrealistically high mobility at moderate vertical electric fields, due to incorrect modelling of interface roughness scattering. The mobility model that was able to offer the closest agreement to reality, whilst including the effects of quantum confinement, also included mobility degradation effects due to the lateral field (HP mobility model). The simulations were therefore somewhat inefficient, but this was a necessary sacrifice in the interests of maximum accuracy.

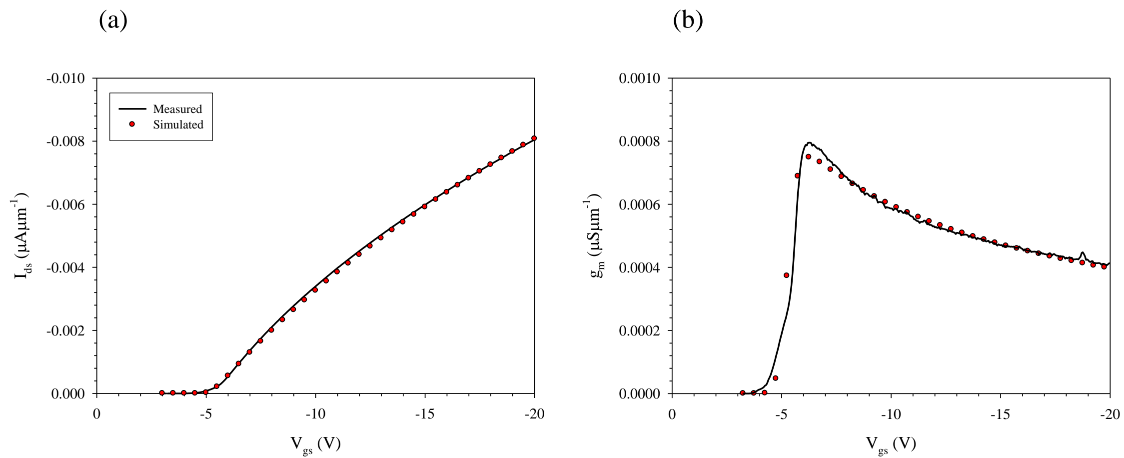


Figure 6.2. Comparison of simulated and measured (a) drain current and (b) transconductance for a 300μm silicon pMOSFET on wafer 10.

Figure 6.2 serves to show how even an apparently very small difference between simulated and measured drain current data is amplified by differentiating to obtain the transconductance, as seen by the slight discrepancy in peak g_m.

6.3.2 Simulation of Strained Silicon pMOSFETs

Having confirmed the validity of the simulation method, strained silicon devices were simulated. Because of the heavy parasitic conduction encountered in the wafers with high germanium content it was anticipated that, using the device parameters extracted in sections 5.3 and 5.4, it would only be possible to achieve good agreement between simulated and measured device characteristics by adjusting the valence band offset between the strained silicon layer and the relaxed SiGe buffer.

An added difficulty in simulating strained silicon devices was the potential for variation in the strained silicon layer thickness across the wafer, in addition to the variation in gate oxide thickness. Whilst the problem of oxide thickness variation was again tackled by conducting quasi-static C-V measurements on the device being measured, it was more difficult to accurately gauge the strained silicon thickness. For this, the X-TEM images from a chip close to that being measured were relied upon. The C-V measurement was also able to provide some confirmation due to the formation of a plateau in the inversion region (section 5.3), though it was not as accurate as the X-TEM micrograph. A similar plateau is also seen in fully pseudomorphic structures with a compressively strained SiGe layer and has been used to extract the valence band offset in this case by comparison with a 1D Poisson Schrödinger solver.[3] In principal it is possible to perform a similar extraction for the case of strained silicon on a relaxed SiGe buffer. However, the large voltages

required in the characterisation of devices from batch k2295 were beyond the capability of such software. Additionally, the C-V measurements, particularly for the lower germanium content wafers, did not provide such an obvious separation of the population of the two layers.

When simulating the strained silicon devices, the first iterations focussed on perfecting the electrical characteristics for low vertical fields, when only the SiGe buffer was occupied. Fortunately the model for relaxed SiGe included in Medici is likely to be accurate, based as it is on the findings of Iyer *et al.*.[4] Apart from a standard correction to the density of states, the only parameters that required slight adjustment were the background doping and the hole mobility, in the same way as for the silicon control. With good agreement between the threshold voltage and first transconductance peak thus achieved, the parameters of the strained silicon layer could be adjusted. As a starting point, this layer was assumed to have the same properties as unstrained silicon. Valence band splitting was simulated by setting the density of states as $4 \times 10^{18}cm^{-3}$ [5] and the hole effective mass was taken as $0.55m_0$, where m_0 is the electronic rest mass.[6] Increasing the carrier mobility allowed good agreement between the height of the measured and simulated second transconductance peak, whilst adjustment of the valence band offset shifted the position of this peak.

Some additional complications also had to be considered. In the absence of any data to the contrary, the relative permittivity of the strained silicon layer was assumed to be the same as that of bulk silicon, at 11.9. In the vertical direction, the tensile strained silicon considered here will be compressed, which may increase the relative

permittivity. To estimate the importance of the accuracy of this parameter, simulations were performed with different values. Arbitrarily increasing the permittivity to 12.5 required the valence band offset to be reduced by ~ 20meV to maintain good agreement between the measured and simulated characteristics. The simulation was also found to be sensitive to the effects of quantum confinement. Neglecting this phenomenon when the strained silicon layer was thin resulted in an unrealistically high occupation of the layer, with the result that the valence band offset had to be increased in order to compensate. In the worst case, which was the device on wafer 2 with a 6nm strained silicon layer, neglecting quantum confinement resulted in an offset ~ 30meV larger. To improve accuracy as far as possible, Medici's pseudo-confinement was calibrated by comparison with the 1D Poisson-Schrödinger solver, Snider,[7] as far as the limits of this simulation would allow.

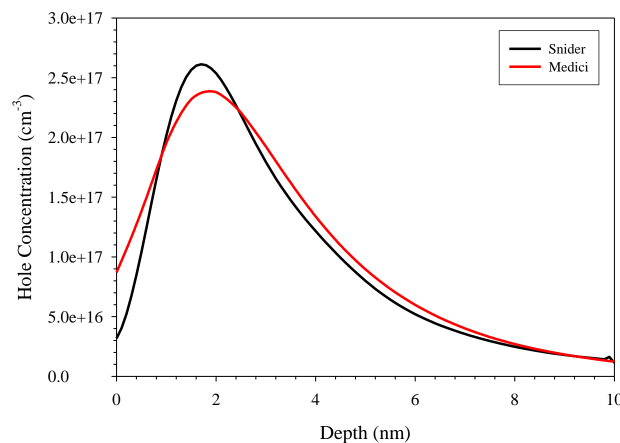


Figure 6.3. Comparison of carrier distribution generated by Snider and Medici for a silicon pMOSFET.

Also of concern was the fact that in reality, there was no sharp division between the strained silicon layer and the SiGe buffer, as would typically be assumed during simulation. Instead there was a slight smearing of the profile, brought about by

segregation of germanium during growth [8] and subsequent diffusion during annealing. This effect was simulated by reducing the thickness of the strained silicon layer and introducing a thin gradeback layer, with linearly varying composition and energy gap. Because occupation of this layer occurred at the interface with the SiGe buffer (figure 6.4), the mobility was taken to be the same as that of the alloy. It was found that the addition of this layer had a very small effect on the simulations, as the hole concentration profile was only slightly altered. Including the layer required a reduction in the valence band offset of less than 5meV, however it was included in all simulations.

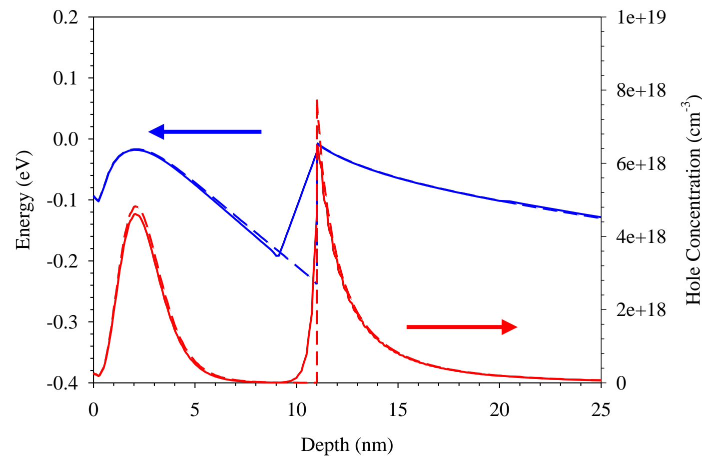


Figure 6.4. Comparison of simulated valence band and hole concentration with (solid lines) and without (dashed lines) a gradeback layer present.

It may be seen (figure 6.5) that the extracted mobility initially matches that of the relaxed SiGe layer very closely but approaches the strained silicon mobility as this layer begins to populate. However, the two mobilities do not coincide as the vertical field continues to increase in strength, owing in part to the fact that the simulated strained silicon mobility does not have the correct field dependence and does not degrade quickly enough.

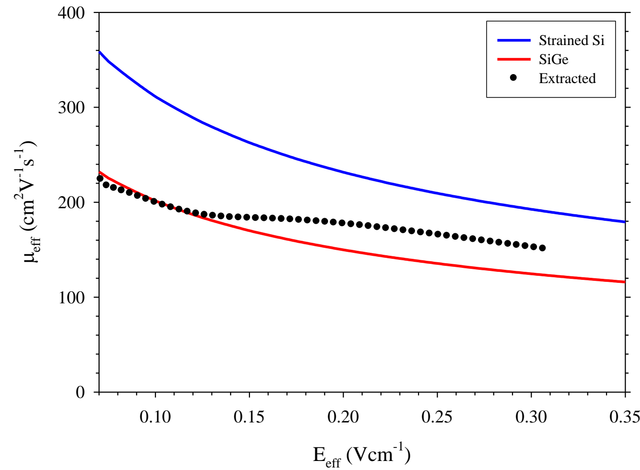


Figure 6.5. Carrier mobility of the two simulated layers, together with the extracted mobility of the real device on wafer 1.

When simulating the strained silicon devices, the fact that there were two wafers for each composition of germanium allowed some confirmation of the extracted value of the valence band offset. In each case, it was the wafer with the thinner strained silicon channel that was simulated first, as this was the one with the greater degree of parasitic conduction, hence providing the most accuracy. The wafer with the thicker channel was then simulated, without further adjustment of the valence band offset. Good agreement was reached in all cases, indicating the accuracy of the method. No simulation was possible for wafers 7 and 8 ($Si_{0.9}Ge_{0.1}$ buffers) because the small valence band offset and thick strained silicon channels resulted in the complete loss of two distinct transconductance peaks.

6.3.3 Results

Figures 6.6 to 6.11 show comparisons of simulated and measured drain current and transconductance for wafers 1 - 6. The valence band discontinuities extracted as a result of these simulations are listed in table 6.1. Errors have been estimated from the

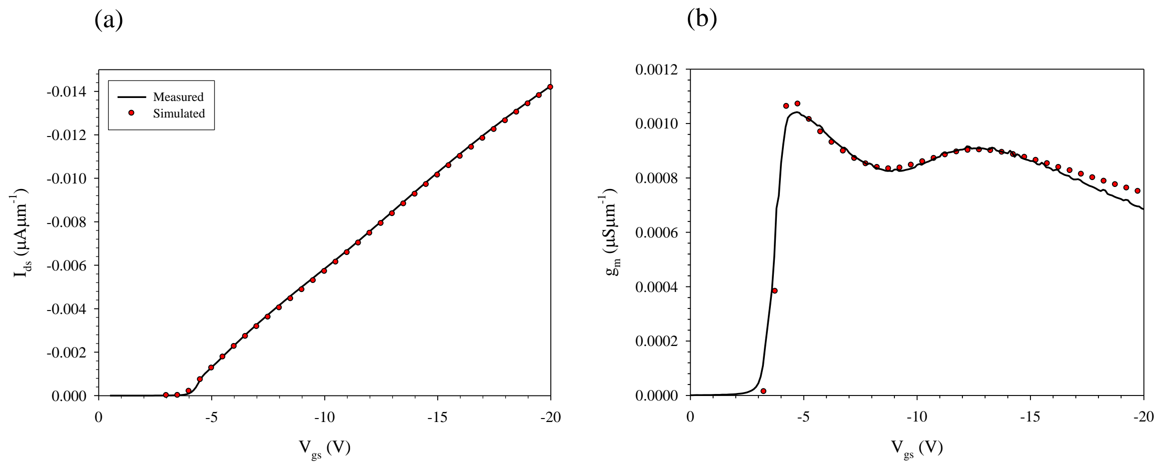


Figure 6.6. Measured and simulated (a) drain current and (b) transconductance for the device on wafer 1.

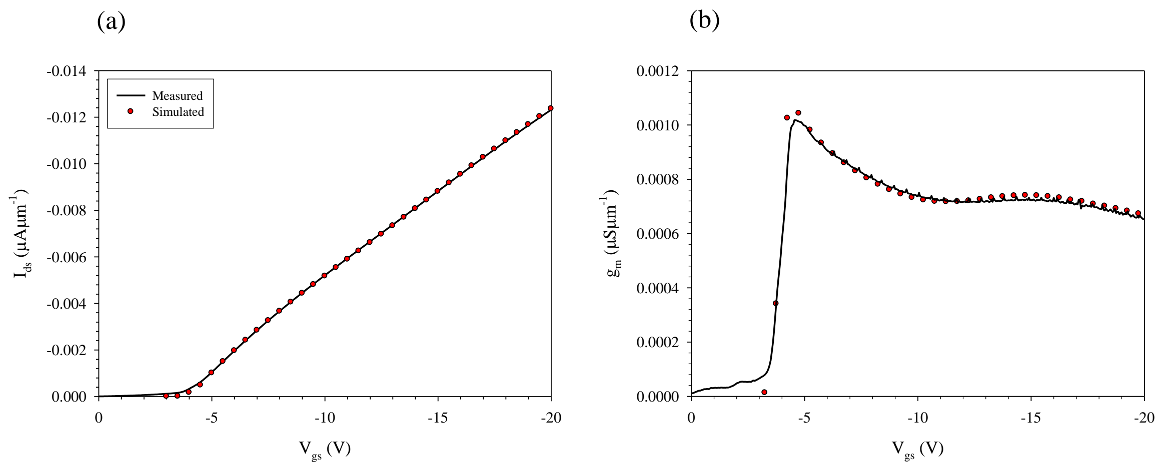


Figure 6.7. Measured and simulated (a) drain current and (b) transconductance for the device on wafer 2.

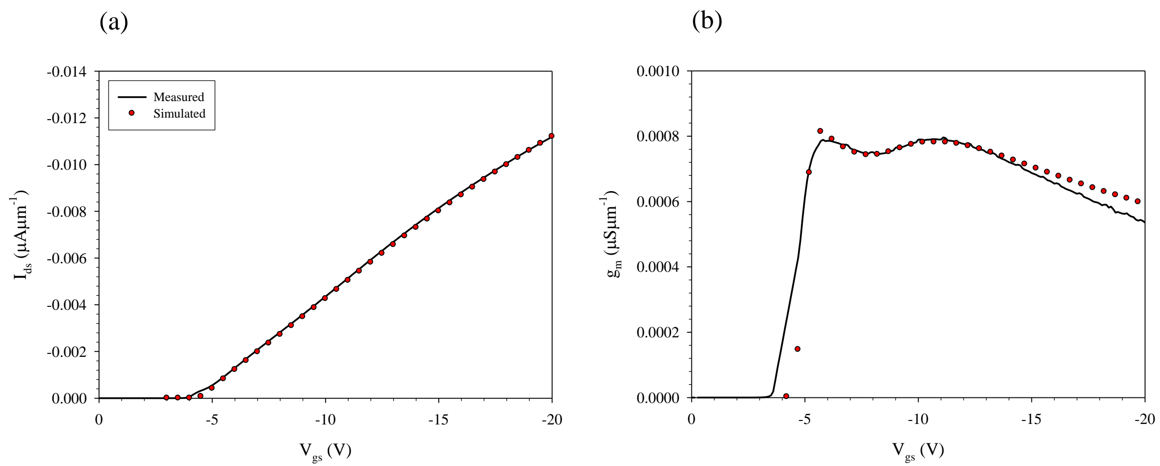


Figure 6.8. Measured and simulated (a) drain current and (b) transconductance for the device on wafer 3.

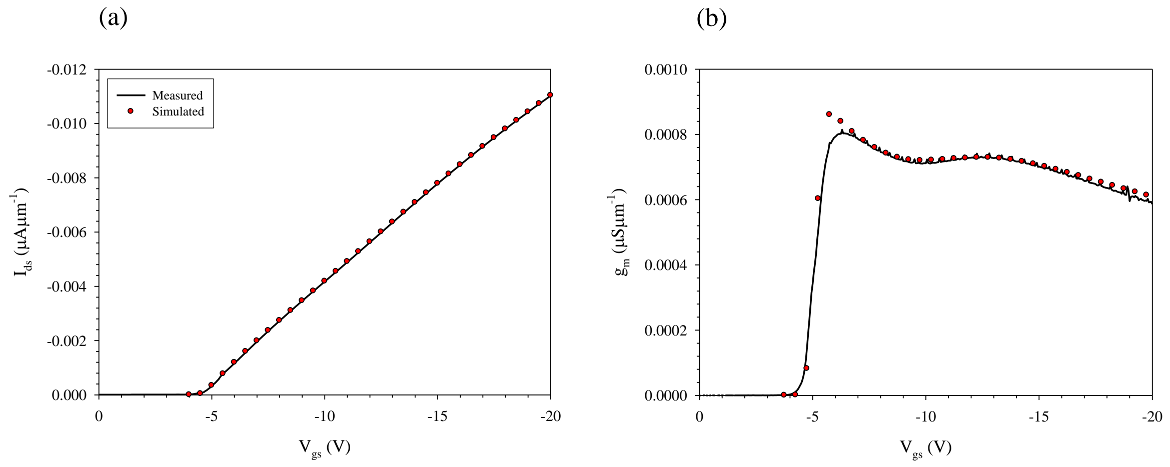


Figure 6.9. Measured and simulated (a) drain current and (b) transconductance for the device on wafer 4.

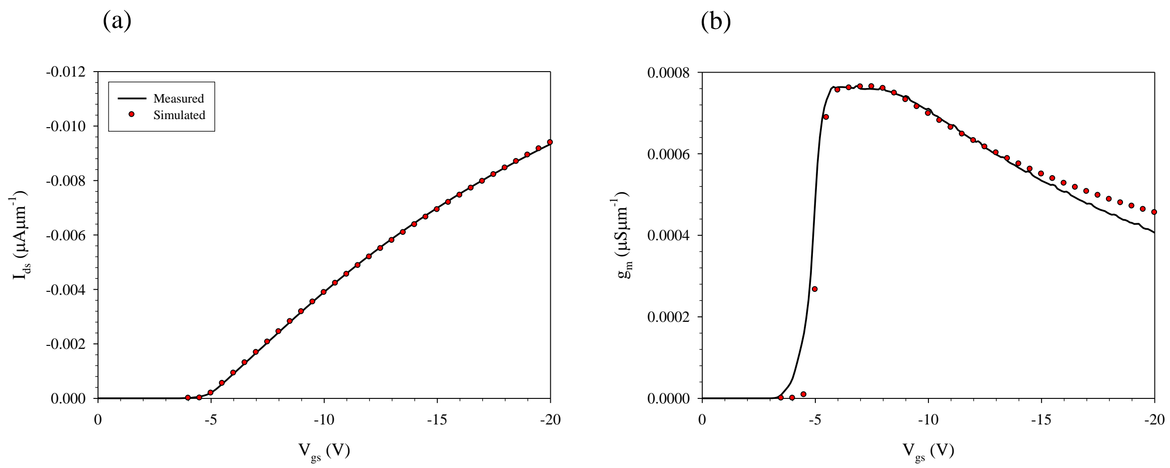


Figure 6.10. Measured and simulated (a) drain current and (b) transconductance for the device on wafer 5.

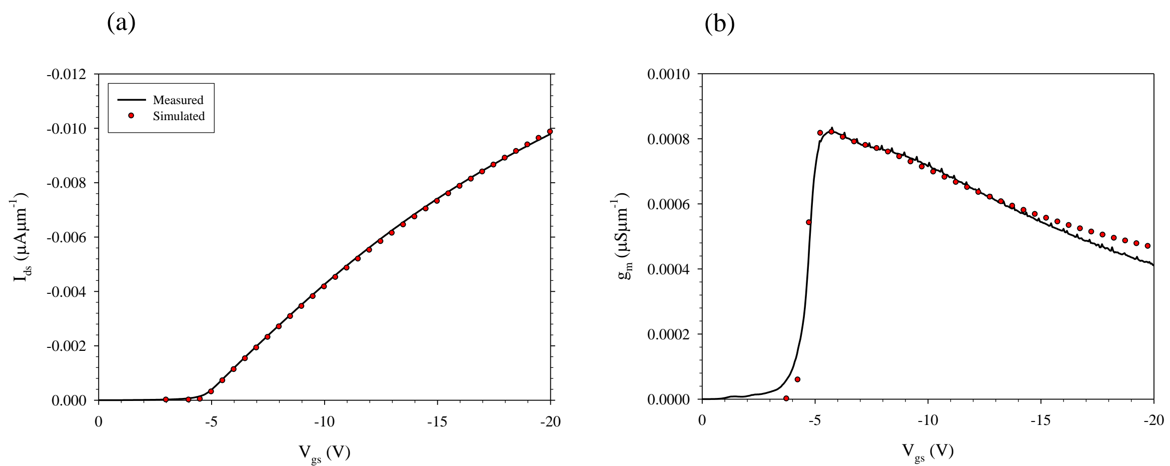


Figure 6.11. Measured and simulated (a) drain current and (b) transconductance for the device on wafer 6.

impact of altering the strained silicon layer thickness by 1nm and accounting for the other uncertainties, such as the unknown relative permittivity of strained silicon. The errors are naturally larger as the germanium content is reduced because the separation of the transconductance peaks became smaller, resulting in a greater range of offsets giving an acceptable fit to the measured data.

Buffer % Ge	36.2	26.6	17.6
ΔE_v (eV)	0.23 ± 0.04	0.19 ± 0.04	0.15 ± 0.05

Table 6.1. Extracted valence band offsets for different mean germanium compositions at 300K.

Maiti *et al.* (9) give the valence band offset for $Si/Si_{0.7}Ge_{0.3}$ as 180meV in their review paper, although they do not reveal their source. The extracted value of 190 ± 40meV for a 27% buffer is in good agreement with this.

6.4 Comparison with Theory

Medici uses a simplified model of the valence bands in a semiconductor and represents the light hole, heavy hole and spin orbit bands as a single energy level. The valence band discontinuities extracted in section 6.3 therefore represent the offset between the weighted average of the strained silicon and the SiGe valence band positions. Rieger and Vogl's theoretical paper (10) considers the bandgaps and band alignments for the general case of a strained $Si_{1-x}Ge_x$ layer on a relaxed $Si_{1-y}Ge_y$ buffer. The weighted average valence band offset (11) is based on original calculations by Van de Walle and Martin (12) and is given by:

$$\Delta E_v^{av} = (0.47 - 0.06y)(x - y). \qquad (6.5)$$

Their predicted results for the wafers considered here are given in table 6.2 and plotted in figure 6.13.

Buffer Ge %	ΔE_v^{av} (Rieger and Vogl) (eV)	ΔE_v^{av} (this work) (eV)
36	0.16	0.23 ± 0.04
27	0.12	0.19 ± 0.04
18	0.08	0.15 ± 0.05

Table 6.2. Comparison of the average valence band offset, ΔE_v^{av}, between strained silicon and relaxed SiGe as predicted by Rieger and Vogl and as extracted in this work.

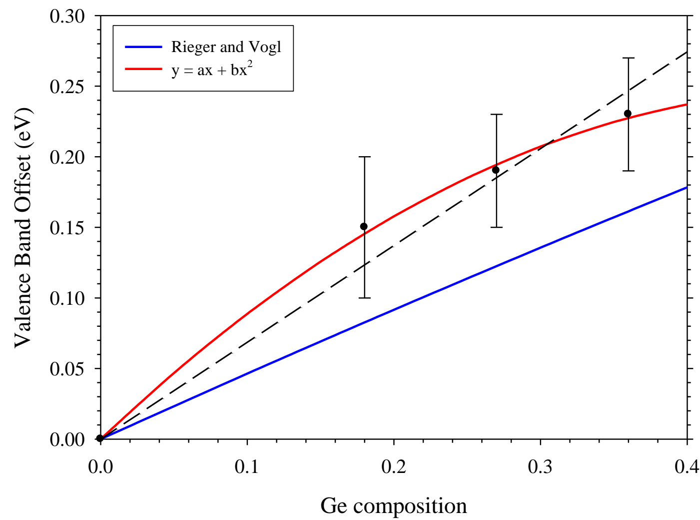


Figure 6.12. Comparison of the extracted values of ΔE_v^{av} with the theoretical predictions of Rieger and Vogl.

It is apparent that ΔE_v^{av} as extracted in this work is somewhat larger than the theoretical predictions. Rieger and Vogl's expression gives a weakly quadratic dependence between the valence band offset and the composition of germanium. A quadratic best fit to the values found in this work is therefore plotted in figure 6.12

(red line). Whilst this provides good agreement with the extracted values, it should be noted that the fit goes on to predict a reducing valence band offset with further increasing germanium content. However, there is sufficient uncertainty in the extracted values that even a linear regression provides a reasonable fit, suggesting than an alternative quadratic dependence may be correct.

A possible explanation for the discrepancy between experiment and theory is that the weighted average valence band offset between strained silicon and relaxed SiGe may change with vertical effective field. As mentioned in section 5.4, it is thought that the effect of the confining potential is to reduce the separation of the light and heavy hole bands in strained silicon, increasing the population of the heavy hole band. Furthermore, there will be some separation of the (normally degenerate) bands in SiGe. Because the strained silicon layers in the devices typically did not populate until $E_{eff} \approx 0.2\text{MVcm}^{-1}$, a somewhat larger valence band offset may have been necessary than would be the case at zero field.

Figure 6.13 shows the familiar type II band alignment of biaxial tensile strained silicon on a relaxed SiGe alloy. It is clear from this diagram that:

$$E_{gSS} + \Delta E_c = E_{gSiGe} + \Delta E_v \,, \tag{6.6}$$

where E_{gSS} and E_{gSiGe} are the bandgaps of strained silicon and relaxed SiGe respectively and ΔE_c and ΔE_v are the conduction and valence band discontinuities between the uppermost energy levels.

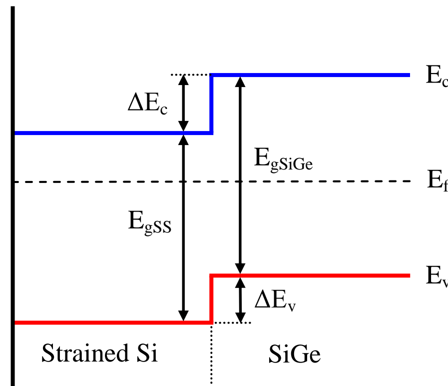


Figure 6.13. The type II band alignment between tensile strained silicon and relaxed SiGe.

It should be noted that whilst the theoretical predictions are self-consistent in that they satisfy equation 6.6, Rieger and Vogl admit that their results for E_{gSiGe} may be suspect. Indeed, comparison of their values with those found by Braunstein *et al.* [(13)] reveals a substantial difference. Given that Braunstein's E_{gSiGe} data was found by optical absorption, it is likely to be the more reliable. The shape of their energy gap versus composition curve is also confirmed by Lang *et al.*,[(14)] although this work is at 90K. Given that ΔE_c for strained silicon on a 33% buffer has been experimentally determined by Garchery *et al.* [(15)] as 290 ± 30meV and agrees well with Rieger and Vogl's prediction, it is possible that the error in E_{SiGe} has been passed on to the valence band offset. In any case, Schäffler [(11)] contends that the approximations and interpolations involved in the theoretical work leave an uncertainty of 100meV. Considering this, the agreement between this work and theory is acceptable.

References

1. Medici 2D Device Simulation Program, Technology Modeling Associates: Sunnyvale, CA, 1994.
2. S. M. Sze, "VLSI Technology", 2nd edition, Bell Telephone Laboratories, pp. 335-336, 1988.

3. S. P. Voinigescu, K. Iniewski, R. Lisak, C. A. T. Salama, J. -P. Noél and D. C. Houghton, "New Technique for the Characterization of Si/SiGe Layers using Heterostructure MOS Capacitors", *Solid-State Electron.*, vol. 37, no. 8, pp. 1491-1501, 1994.
4. S. S. Iyer, G. L. Patton, J. M. C. Stork, B. S. Meyerson and D. L. Harame, "Heterojunction Bipolar Transistors Using Si-Ge Alloys", *IEEE Trans. Elec. Dev.*, vol. 36, no. 10, pp. 2043-2064, 1989.
5. L. Yang *et al.*, "Si/SiGe Heterostructure Parameters for Device Simulation", *unpublished.*
6. L. S. Geux and K. Yamaguchi, "Modeling and Characterization of a Strained Si/$Si_{1-x}Ge_x$ Transistor with δ-Doped Layers", *J. Appl. Phys.*, vol. 86, no. 3, pp. 1443-1448, 1999.
7. 1D Poisson: G. Snider, University of Notre Dame, IN, 2001.
8. S. Fukatsu, K. Fujita, H. Yaguchi, Y. Shiraki and R. Ito, "Self-Limitation in the Surface Segregation of Ge Atoms during Si Molecular Beam Epitaxial Growth", *Appl. Phys. Lett.*, vol. 59, no. 17, pp. 2103-2105, 1991.
9. C. K. Maiti, L. K. Bera and S. Chattopadhyay, "Strained-Si Heterostructure Field Effect Transistors", *Semicond. Sci. Technol.*, vol. 13, pp. 1225-1246, 1998.
10. M. M. Rieger and P. Vogl, "Electronic-Band Parameters in Strained $Si_{1-x}Ge_x$ Alloys on $Si_{1-y}Ge_y$ Substrates", *Phys. Rev. B*, vol. 48, no. 19, pp. 14276-14287, 1993.
11. F. Schäffler, "High-mobility Si and Ge structures", *Semicond. Sci. Technol.*, vol. 12, pp. 1515-1549, 1997.
12. C. Van de Walle and R. M. Martin, "Theoretical Calculations of Heterojunction Discontinuities in the Si/Ge System", *Phys. Rev. B*, vol. 34, no. 8, pp. 5621-5634, 1986.
13. R. Braunstein, A. R. Moore and F. Herman, "Intrinsic Optical Absorption in Germanium-Silicon Alloys", *Phys. Rev.*, vol. 109, no. 3, pp. 695-710, 1958.
14. D. V. Lang, R. People, J. C. Bean and A. M. Sergent, "Measurement of the Band Gap of Ge_xSi_{1-x}/Si Strained-Layer Heterostructures", *Appl. Phys. Lett.*, vol. 47, no. 12, pp. 1333-1335, 1985.
15. L. Garchery, I. Sagnes, P. A. Badoz, "Conduction Band Discontinuity and Electron Mobility in a Strained Si/SiGe Heterostructure", *Applied Surface Science*, vol. 102, pp. 202-207, 1996.

Chapter 7

Conclusion

7.1 Summary

A batch of long strained silicon pMOSFETs on relatively high germanium composition virtual substrates has been thoroughly characterised (k2295). It was found that at these geometries, strained silicon is able to offer an improvement in mobility, and hence drain current, over conventional silicon pMOS. The enhancements were smaller than expected due to extensive parasitic conduction and poor material quality. The enhancement was reduced with the channel length, which has been attributed to problems with the source and drain implantation. In addition, at high drain bias, velocity saturation and self-heating removed the advantage offered by strained silicon altogether. A side-effect of the high threading dislocation density (in excess of $10^{8}cm^{-2}$ for the highest germanium contents) was very high off-state leakage current. The uniformity of the batch was poor. For the strained silicon wafers this appeared to be due to a combination of growth and fabrication problems.

Device simulation of the characterised silicon and strained silicon pMOSFETs has allowed the valence band discontinuity between biaxial tensile strained silicon and relaxed SiGe to be extracted, by comparison of transconductance characteristics. To the best of the author's knowledge, this is the first time that a determination of the valence band offset has been based on experiment, rather than solely theoretical considerations. The discontinuity was found to be somewhat larger than previous theoretical predictions, although there are sufficient uncertainties surrounding both sets of data for the discrepancy to be explained. This work is to be published.

A novel pulsed measurement system has been constructed and used to measure strained silicon MOSFETs in the absence of self-heating. Appreciable self-heating was observed in both pMOSFETs and nMOSFETs. It was found that a pulse duration of 125ns was insufficient to completely remove the heating effect, whereas pulses of 10 and 25ns were found sufficient. By performing pulsed measurements at elevated temperatures, the thermal resistance of strained silicon MOSFETs was extracted. These were found to be lower than expected, which was attributed to the existence of an additional thermal conduction path. The feasibility of introducing strained silicon on virtual substrates to the current technology node from a thermal standpoint has been considered.

Additional measurements have been made on a previously characterised strained silicon wafer, found to exhibit very high hole mobility. The mobility enhancement has been confirmed and an investigation of carrier injection velocity has shown that strained silicon is not operating near the thermal limit for channel lengths down to 100nm, even in the presence of extensive DIBL. This work is to be published. It has also been found that impact ionisation for holes is apparently reduced in tensile strained silicon compared to unstrained silicon.

7.2 Discussion and Suggestions for Further Work

Tensile strained silicon appears to offer considerable advantages over conventional silicon CMOS. Even with substantial parasitic conduction, strained silicon pMOSFETs showed a highly significant drive current enhancement over silicon pMOS at long channel lengths. This enhancement is expected to diminish with decreasing channel length due to velocity saturation but, with such a large mobility

enhancement, should remain significant. Unfortunately, in these investigations, self-heating and apparent complications with device processing (k2295) or a poor control (k2334) prevented a fair comparison of drive current between strained and unstrained silicon pMOSFETs at short channel lengths. The problems that caused degradation of the strained silicon low field mobility enhancement as the channel length of the devices in batch k2295 was reduced appear to be an unfortunate feature of the batch, rather than some fundamental problem with strained silicon, as no other authors have reported such an effect. Andrieu *et al.* [(1)] report a similar phenomenon, but as their devices have a compressively strained SiGe channel rather than strained silicon, this lends weight to the theory that it is a processing issue.

Of great relevance to aggressively scaled devices is the electron and hole mobility enhancement at vertical fields of approximately $1MVcm^{-1}$. The obvious extension to this work is to characterise devices at low temperature. By removing much of the phonon scattering, it should be possible to ascertain whether interface roughness at the silicon/oxide interface is indeed reduced by the addition of strain. Monte Carlo simulations appear to indicate that this must be the case, in order to explain the continuing electron mobility enhancement observed by other authors at high vertical fields.[(2-4)] A similar explanation may also be required to explain the predicted 25% hole mobility enhancement over the universal curve for batch k2334, at a vertical field of $1MVcm^{-1}$. Previously, hole mobility enhancements in tensile strained silicon have been observed to vanish at high fields,[(5)] so this result is highly significant.

It is apparent that a major obstacle preventing the successful introduction of tensile strained silicon into mainstream manufacturing is the virtual substrate. The self-

heating effects on a typical 2μm thick virtual substrate have been shown to be problematic, and this work has shown that strained silicon pMOSFETs are also likely to suffer in this regard. Recently, Polonsky and Jenkins [6] have shown that the temperature of a strained silicon MOSFET rises and falls exponentially when it is turned on and off, both with a thermal time constant of 62ns. This work agrees well with that finding, as 125ns pulses were not sufficient to completely remove self-heating, whereas the 10 and 25ns pulses employed using the fast generator overcame the problem completely. It may be possible to add processing steps to the fabrication of strained silicon MOSFETs, to introduce additional paths for thermal conduction in the regions that require it. The cooling apparently offered by using a slightly unusual device design in this work indicates the feasibility of this approach. A better (and cheaper) solution however, would be to substantially reduce the thickness of the virtual substrate. This of course has severe implications for the other Achilles heel of the strained Si/SiGe heterostructure: dislocations.

Threading dislocations appear to have a crippling effect on the standby power of strained silicon devices, a consideration that would be a major problem in industry, where the ever increasing power consumption of CMOS chips is a real concern. It now seems likely that if tensile strained silicon is introduced, it may be in the form of strained silicon on insulator (SSOI).[7], [8] By this technique, it may be possible to remove the SiGe layer altogether, by transferring only the strained silicon onto oxide. Of course, the strained silicon is still originally created by growth on a virtual substrate, however in this case it may be made as thick as required to reduce the dislocation density. Self-heating problems, due to the oxide, will of course remain.

Intel® have already introduced strained silicon to their fabrication process, although in this case no virtual substrate is involved. Instead, uniaxial strain is mechanically created using a nitride capping film to introduce tensile strain to the nMOSFETs, and using selective epitaxial deposition of $Si_{0.83}Ge_{0.17}$ in the source/drain regions to introduce compressive strain to the pMOSFETs. Using compressive strain, Thompson *et al.* (9) are able to demonstrate that their hole mobility enhancement of 50% is largely undiminished by a $1MVcm^{-1}$ vertical field. However, the degree of tensile strain that is imparted by the nitride film remains unclear, and it may yet be necessary to utilise virtual substrates in order to realise the full electron mobility enhancement.

Tensile strained silicon also appears to offer reduced impact ionisation characteristics, at least for holes, which is of relevance for high frequency power amplifier functions. It is unclear why strained silicon should offer any reduction in impact ionisation, and the results of this investigation are surprising in light of the fact that Waldron *et al.* (10) found elevated impact ionisation for electrons in strained silicon. The fact that their test structure included conduction through the relaxed SiGe could have had a large impact on their results, and it is unclear what proportion of electrons was confined in each layer. Of course, self-heating is a major issue to be overcome in high power applications. This area certainly merits further work.

The discovery that the valence band discontinuity between tensile strained silicon and relaxed SiGe may be larger than theoretical calculations had supposed is an important result from the point of view of strained silicon pMOSFET design. No definitive information on this parameter exists and experimentalists have often used different

values in their work.[11], [12] Although the offsets extracted in this work are average offsets, and are also affected by the presence of a vertical effective field, these parameters provide a useful guideline for future Medici simulations of strained silicon devices.

Batch k2295 turned out to be well suited to a determination of valence band offset, with its relatively high germanium contents and extremely thick gate oxides. However, had it been specifically designed for such work, then the strained silicon layers could have been specified to be much thinner. It may have been possible to extract the offset for the 10% buffers with strained silicon layer thicknesses of a few nanometres, instead of 25 and 40nm. The accuracy of the method for the other wafers could also have been thus improved. Further work in this area would therefore involve a greater range of germanium compositions, combined with thinner strained silicon layers. Another possibility is the use of more sophisticated simulation software which models the different subbands, rather than using a single, averaged band. Such a simulation would allow direct comparison of the offset between the uppermost energy levels of strained silicon and relaxed SiGe with that predicted by Rieger and Vogl.[13]

References

1. F. Andrieu *et al.*, "SiGe Channel p-MOSFETs Scaling-Down", *33rd European Solid-State Device Research Conference Proceedings*, pp. 267-270, 2003.
2. M. V. Fischetti, F. Gámiz and W. Hänsch, "On the Enhanced Electron Mobility in Strained-Silicon Inversion Layers", *J. Appl. Phys.*, vol. 92, no. 12, pp. 7320-7324, 2002.
3. L. Yang, J. R. Watling, R. C. W. Wilkins, J. R. Barker and A. Asenov, "Reduced Interface roughness in Sub-100nm Strained Si *n*-MOSFETs - A Monte Carlo Simulation Study", *5th European Workshop on the Ultimate Integration of Silicon*, pp. 23-26, 2004.

4. J. R. Watling *et al.*, “The Impact of Interface Roughness Scattering and Degeneracy in Relaxed and Strained Si n-channel MOSFETs”, Solid-State Electon., vol. 48, no.8, pp. 1337-1346, 2004.
5. K. Rim *et al.*, “Strained Si CMOS (SS CMOS) Technology: Opportunities and Challenges”, *Solid-State Electron.*, vol. 47, no. 7, pp. 1133-1139, 2003.
6. S. Polonsky and K. A. Jenkins, “Time-Resolved Measurements of Self-Heating in SOI and Strained-Silicon MOSFETs Using Photon Emission Microscopy”, *IEEE Elec. Dev. Lett.*, vol. 25, no. 4, pp. 208-210, 2004.
7. Z. Cheng *et al.*, “Fully Depleted Strained-SOI n- and p-MOSFETs on Bonded SGOI Substrates and Study of the SiGe/BOX Interface”, *IEEE Elec. Dev. Lett.*, vol. 25, no. 3, pp. 147-149, 2004.
8. K. Kim, C.-T. Chuang, K. Rim and R. V. Joshi, “Performance Assessment of Strained-Si Channel-on-Insulator (SSOI) CMOS”, *Solid-State Electron.*, vol. 48, no. 2, pp. 239-243, 2004.
9. S. E. Thompson *et al.*, “A Logic Nanotechnology Featuring Strained-Silicon”, *IEEE Elec. Dev. Lett.*, vol. 25, no. 4, pp. 191-193, 2004.
10. N. S. Waldron *et al.*, “Impact Ionization in Strained-Si/SiGe Heterostructures”, *IEEE IEDM Tech. Dig.*, 2003.
11. K. Rim, J. Welser, J. L. Hoyt and J. F. Gibbons, “Enhanced Hole Mobilities in Surface-Channel Strained-Si p-MOSFETs”, *IEEE IEDM Tech. Dig.*, 1995.
12. C. K. Maiti, L. K. Bera and S. Chattopadhyay, “Strained-Si, Heterostructure Field Effect Transistors”, *Semicond. Sci. Technol.*, vol. 13, pp. 1225-1246, 1998.
13. M. M. Rieger and P. Vogl, “Electronic-Band Parameters in Strained $Si_{1-x}Ge_x$ Alloys on $Si_{1-y}Ge_y$ Substrates”, *Phys. Rev. B*, vol. 48, no. 19, pp. 14276-14287, 1993.