The Development of Sub-25 nm III-V High Electron Mobility Transistors

A THESIS SUBMITTED TO
THE DEPARTMENT OF ELECTRONICS AND ELECTRICAL ENGINEERING
FACULTY OF ENGINEERING
UNIVERSITY OF GLASGOW
IN FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

By
Steven Bentley
July 2009

“Le bon Dieu est dans le détail.”

Gustave Flaubert

“Der Teufel steckt im Detail.”

Blixa Bargeld

Abstract

High Electron Mobility Transistors (HEMTs) are crucially important devices in microwave circuit applications. As the technology has matured, new applications have arisen, particularly at millimetre-wave and sub-millimetre wave frequencies. There now exists great demand for low-visibility, security and medical imaging in addition to telecommunications applications operating at frequencies well above 100 GHz.

These new applications have driven demand for high frequency, low noise device operation; key areas in which HEMTs excel. As a consequence, there is growing incentive to explore the ultimate performance available from such devices.

As with all FETs, the key to HEMT performance optimisation is the reduction of gate length, whilst optimally scaling the rest of the device and minimising parasitic extrinsic influences on device performance.

Although HEMTs have been under development for many years, key performance metrics have latterly slowed in their evolution, largely due to the difficulty of fabricating devices at increasingly nanometric gate lengths and maintaining satisfactory scaling and device performance. At Glasgow, the world-leading 50 nm HEMT process developed in 2003 had not since been improved in the intervening five years.

This work describes the fabrication of sub-25 nm HEMTs in a robust and repeatable manner by the use of advanced processing techniques: in particular, electron beam lithography and reactive ion etching. This thesis describes firstly the development of robust gate lithography for sub-25 nm patterning, and its incorporation into a complete device process flow. Secondly, processes and techniques for the optimisation of the complete device are described.

This work has led to the successful fabrication of functional 22 nm HEMTs and the development of 10 nm scale gate pattern transfer: simultaneously some of the shortest gate length devices reported and amongst the smallest scale structures ever lithographically defined on III-V substrates. The first successful fabrication of implant-isolated planar high-indium HEMTs is also reported amongst other novel secondary processes.

Acknowledgements

I am indebted to many people for their input on this work, without whom it would never have been completed. I would like to thank Iain Thayne for his unique mentoring input, and for striking a great balance as supervisor between allowing me freedom and providing guidance when required. He also provided many curries.

All the members of the Ultrafast Systems Group both past and present deserve thanks for their transfer of knowledge and sharing in the many delights of fabrication. Particular thanks are due to Dave Moran and Richard Hill for their effective concoction of camaraderie, wisdom and sarcasm (1:1:1), and for so many thought-provoking discussions. I am indebted to Xu Li for developing the dry etch processes used in this work, and for his apparently telepathic control over SF_6 etch times, and to Helen McLelland and Susan Ferguson for their support and optimistic understanding. Kevin Docherty is to be praised for his encyclopaedic familiarity with the wiles of the VB6 and for his essential guidance on alignment techniques. Thanks are also due to Billy Smith and Colin How for their skills in TEM sample preparation and imaging.

The entire staff of the James Watt Nanofabrication Centre deserves thanks for the smooth running of the facilities, which is a testament to their dedication and hard work.

Heartfelt thanks are also due to my family and friends, who have supported me through my studies, and whose unwavering belief and patience have enabled me to remain largely sane. My mother deserves a medal for her many hours of treating physics as a foreign language whilst proof-reading, which have transformed this thesis into something much less unintelligible. Undertaking this work has been made possible by the constant love and encouragement of Vicki Allan, who has seen the incoherent, obsessive-compulsive side and decided to marry me anyway.

Finally, to all those friends and colleagues who are too many to name, but who have made such a huge difference, untold thanks are due. The combined humour, kind words, prayers and wisdom have transformed this work from a worthwhile academic pursuit to an enjoyable life experience. Many thanks.

Contents

List of Tables

List of Figures

1. Associated Publications

- S. Bentley, X. Li, D.A.J. Moran and I.G. Thayne, "Two methods of realising 10 nm T-gate lithography", *Microelectronic Engineering* (2009) vol. 86 pp. 1067-1070
- S. Bentley, X. Li, D.A.J. Moran and I.G. Thayne, "Fabrication of sub-25 nm InGaAs/InAlAs HEMTs by two-step gate lithography", presented at HETECH, Venice, Italy, 2008.
- S. Bentley, R.J.W. Hill, R. Gwilliam and I.G. Thayne, "Implant-isolated InGaAs/InAlAs HEMTs with f_t of 420 GHz", presented at UK Semiconductors, Sheffield, U.K., 2008.
- S. Bentley, X. Li, D.A.J. Moran and I.G. Thayne, "Fabrication of 22 nm T-gates for HEMT applications", *Microelectronic Engineering* (2008) vol. 85 pp. 1375-1378.
- S. Bentley, X. Li, D.A.J. Moran and I.G. Thayne, "Fabrication of 22 nm T-gates for HEMT applications", presented at UK Compound Semiconductors, Sheffield, U.K., 2007.

2. Introduction

Since William Shockley's discovery of the Field Effect in 1945, and the subsequent invention of the Field Effect Transistor in the 1960s, FETs have become the primary component in modern electronics, owing much to their simplicity, efficiency of operation and ease of fabrication.

The genesis of the HEMT in 1979 [1] was in many ways accidental. At the time, Mimura, the inventor of the HEMT, was working on GaAs MOSFET development for high-speed logic, using native oxides as the gate insulator; a system which did not allow inversion or accumulation due to high surface state density. In 1978, Dingle et al., of Bell Labs, NJ, published work on a new method of doping in heterostructures [2], making use of the relatively new field of molecular beam epitaxy (MBE) to separate electrons from the dopants. Until this point, it had been impossible to achieve structures which simultaneously exhibited high electron density and high mobility.

Since Mimura's work on the MOSFET had proved fruitless, he began work on the confinement of electrons in a modulation-doped heterostructure, envisioning a structure incorporating a Schottky metal gate placed on the AlGaAs side of a single AlGaAs/GaAs heterostructure. Both enhancement and depletion-mode devices were demonstrated by mid-1980. The first HEMT logic circuits were subsequently reported in 1981, and the first low-noise amplifiers entered commercial production in 1985.

The HEMT has since found many applications, particularly in microwave communications and low-noise detectors. Though designed for logic, Schottky gate devices suffer from large gate leakage currents when driven in forward bias, and the high cost of III-V wafers when compared to silicon. As a consequence, III-V logic has not yet replaced silicon. The HEMT, however, has a variety of uses as the demand for telecommunications has boomed over the last several decades, and new applications have emerged as the upper end of the frequency spectrum has become accessible. In particular, the HEMT

excels as a low-noise device, and has found myriad uses in imaging and amplification applications. In recent years, the invention of "millimetre-wave" imaging systems has called for high-sensitivity, low noise detectors and circuits at frequencies in excess of 100 GHz. Millimetre-wave operation refers to the wavelength of the radiation of 1-10 mm, corresponding to frequencies of 30-300 GHz, whilst sub-millimetre-wave radiation refers to wavelengths of 0.1-1 mm, covering the spectrum from 300 GHz - 3 THz. Electromagnetic waves of these frequencies can penetrate low-density materials such as fabrics or liquids with ease, but are reflected by higher-density materials such as metals. As a result, this frequency range is of great use in satellite, astronomical, low-visibility, security and medical imaging particularly in addition to its communications applications [3].

It is important to note that the HEMT is not the only candidate device in these fields, though it has unique advantages. Its principal contemporary is the Heterojunction Bipolar transistor (HBT), which for a great many years trailed the HEMT in performance, predominantly due to its current gain dependence on emitter area and limitations in parasitic resistances [4]. In recent years, however, advances in InP HBT technology have led to HBTs competing with and exceeding the high-frequency performance of HEMTs [5], predominantly due to the evolution of the technology towards double heterojunction layouts exploiting pseudomorphic active layers on InP. Presently, an HBT holds the record cutoff frequency of 765 GHz [6].

There are, however, major differences in the two devices. HEMTs intrinsically exhibit lower noise than HBTs, due to capacitive coupling between gate and channel which results in uniquely low noise figures [7]. This is unique to HEMTs and as a result, they generally exhibit lower noise figures than HBTs at high frequency. As a result, when high-frequency noise is a concern, as for many MMIC applications such as receivers, the HEMT will continue to be the device of choice [8, 9].

Fabrication technologies are also fundamentally different, with FETs requiring simultaneous lateral and vertical scaling, whilst bipolar transistors principally require vertical epitaxial scaling, although the highest-frequency HBTs feature sub-micron emitter widths. Arguably, HBT technology may, as a result, continue to be more scalable than that of the more lithographically dependent HEMT in the future [4, 10]. Conversely, HEMT process flows are more similar to conventional CMOS processes and are easily adapted for mass production, and, critically, are unipolar devices, more readily suited to integration with CMOS.

In recent years, much speculation has been made as to direction at the end of the silicon

roadmap below the 22 nm technology generation, with much effort being expended on alternative, high mobility channel materials, in which HEMT-like systems are the most likely candidate. As a result, several groups have focussed research efforts either on the adaptation of III-V HEMTs to digital applications [11–13] or the development of III-V MOSFET technology based on the HEMT [14–17].

For these reasons, HEMTs, already a mature technology, are likely to be the focus of much development for the foreseeable future.

Although HEMTs have been under development for many years, key performance metrics have latterly slowed in their evolution, largely due to the difficulty of fabricating devices at increasingly nanometric gate lengths and maintaining satisfactory scaling and device performance, with many groups halting device fabrication at 50-100 nm gate lengths. In recent years, however, advances in electron beam lithography, plasma processing and molecular beam epitaxy have once more opened up the ability to fabricate aggressively-scaled structures.

This project aims to fabricate well-scaled HEMTs at gate lengths of less than 25 nm. This has required the development of several entirely new process modules and complete revision of traditional HEMT process flows, building on the foundations of earlier work on non-annealed ohmic contacts, gate lithography and double-delta-doped device structures.

This thesis describes first the theory underlying HEMT operation and the processes and methods used to fabricate and characterise the materials and structures required for their realisation. A review of current research will then be presented before extensively describing the processes and techniques developed for reliable short gate-length device fabrication and the devices resulting from this work.

3. The High Electron Mobility Transistor

3.1 Introduction

The High Electron Mobility Transistor (HEMT) is a unipolar field effect device, relying on the modulation of the electron population of a channel between source and drain to control the drive current of the transistor. Where HEMTs differ from MOSFET or MESFET devices, however, is in the use of multiple heterojunctions in the device layer structure, which act to confine electrons, improving transport characteristics. Such heterostructures are formed by the union of two dissimilar semiconductors. This junction of two materials with different bandgaps results in the formation of discontinuities in the structure of the conduction and valence bands through the device, which create the confinement effects that prove so desirable in improving device performance. As a result, HEMTs are also known as HFETs (Heterojunction FET) or MODFETs (Modulation-Doped FET), monikers which, given the non-linear transport effects observed in short-channel devices, described later in this chapter, may be more descriptive than the 'HEMT' acronym by which the devices are historically known.

This chapter will describe the operation of these devices and the underlying physics which governs it. In doing so, the material properties, electron dynamics and issues pertaining to the effective scaling of HEMTs will be discussed, and the figures of merit and equivalent circuits used as metrics for device performance will be outlined.

3.2 Device overview

The HEMT consists of a multi-layer stack of semiconductor materials which comprises the vertical architecture of the device. In the case of the devices forming the scope of this work, this is formed using the InGaAs/InAlAs materials system and typically realised by Molecular Beam Epitaxy (MBE) or Metal-Organic Chemical Vapour Deposition (MOCVD), which will be described in Chapter 4. HEMTs, like all FETs, rely on the application of a voltage between source and drain to create a current flow in a channel region. For a given source-drain voltage, the electron population of this channel, and hence the current flow, is then controlled by the application of a gate voltage.

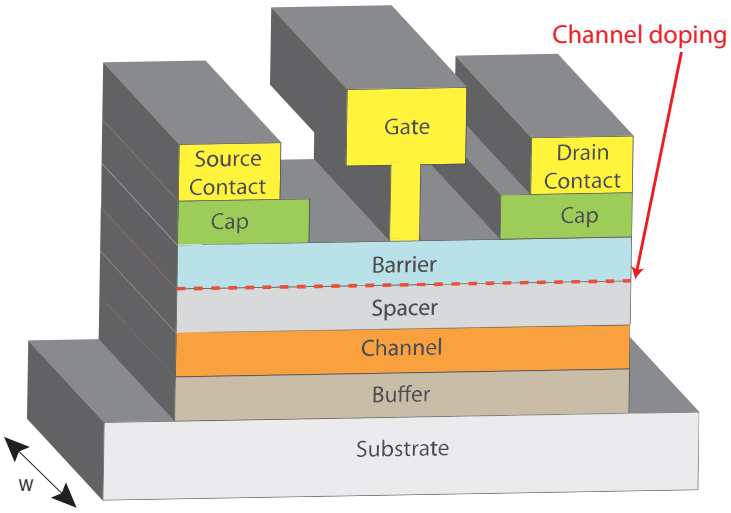


Figure 3.1: General layout of a HEMT.

Figure 3.1 shows a general schematic of a HEMT in cross-section. The channel of the device is typically formed in InGaAs and is buried by several ancillary layers; the spacer, barrier and cap layers, all of which are formed during wafer growth. Doping is also introduced during wafer growth in order to populate the channel, usually in the form of delta doping, described in Section 3.4.1. The cap layer is highly-doped and allows the formation of low-resistance source and drain contacts to the device channel (Section 3.5.4), and is selectively etched away between the source and drain during device fabrication, exposing the undoped barrier layer. The gate is defined on the surface of the barrier layer. The role of the spacer layer is to enhance electron mobility in the channel, as will

be described in Section 3.4.1.

The source and drain contacts are metallic and must be ohmic in nature (Section 3.5.4), whilst the gate is also metallic, but forms a Schottky contact to the barrier, as discussed in Section 3.5.3.

The buffer layer shown in Figure 3.1 has the dual purpose of providing a high-quality surface with a lattice constant similar to that of the channel for optimal channel growth, and to provide electron confinement in the channel, preventing real space transfer and leakage into the buffer.

HEMTs are generally considerably wider than their length and are completed by coupling a thick metallic bondpad layer to the three contacts.

Understanding HEMT operation requires first an understanding of the underlying materials physics, so it is this on which the next sections are focussed.

3.3 Electron transport in semiconductors

Conduction in semiconductors is defined by their band structure - particularly the bandgap of the material defined by the separation of the conduction and valence bands, and resulting from the quantum mechanical properties of matter. Electrons in atoms are constrained to quantised energies resulting from the Schrödinger wave equation associated with a potential well formed around the atom's nucleus [18]. When many atoms are then arranged in solids, these energy levels form bands of closely-spaced energy levels [19]. Since electrons are fermions, states in low energy bands quickly become filled, leaving empty states in upper energy bands. These low-energy bands are known as valence bands, whilst upper, usually empty, states are known as the conduction bands. In metals, the lower conduction band states are partly filled, whilst in intrinsic semiconductors, most conduction band states are empty, rendering the material high-resistivity [20]. The Fermi-Dirac distribution function describes the probability of a state being filled, and the Fermi level, E_f, describes the energy at which this probability is one half [21]. The Fermi energy, given by Equation 3.1, is usually positioned mid-bandgap in intrinsic semiconductors:

$$E_f = \frac{E_c + E_v}{2} + \frac{k_B T}{2} \ln \frac{N_v}{N_c} \tag{3.1}$$

Where E_c and E_v are the conduction and valence band minimum and maximum, respectively, k_B is the Boltzmann constant, T is temperature, and N_c and N_v are the conduction and valence band densities of states.

Electrons may be thermally excited from the valence band to the conduction band, allowing conduction to occur either by the electrons themselves or by vacated holes in the valence band. At room temperature, however, electron thermal energies are generally an order of magnitude lower than the bandgap of all but the narrowest-bandgap semiconductors, and intrinsic semiconductors are highly resistive. Different semiconductors exhibit various bandgaps with accordingly varying properties.

Doping a semiconductor by introducing impurities from adjacent groups of the periodic table introduces further free charge carriers once they become ionised, enhancing conductivity. Such impurities are either classified as donors, where excess electrons are introduced, or acceptors, where an extra hole is created, introducing excess positive charge carriers. Donors introduce extra energy levels at the bottom of the conduction band, shifting the Fermi level upwards, whilst acceptors create levels near the top of the valence band, shifting the Fermi level downwards. Situations where the semiconductor has excess electrons are referred to as n-type, whilst materials where holes are the majority charge carrier are known as p-type. In the case of III-V materials, silicon is a commonly used n-dopant, whilst beryllium or carbon are frequently used to p-dope.

For n-type materials, the Fermi level can be expressed with respect to the density of states in the conduction band, N_c and the donor concentration, N_d:

$$E_f = E_c - k_B T \ln \frac{N_c}{N_d} \tag{3.2}$$

Since electrons are the majority charge carriers in HEMTs, the remaining chapter focusses on electron transport exclusively.

3.3.1 Drift, diffusion, scattering and mobility

Electrons in the conduction band move similarly to electrons in free space, in that their behaviour can be described by the three-dimensional Schrödinger wave function as a consequence of the three-dimensional periodic potential of the crystal [18] :

$$\psi_{\mathbf{k}}(\mathbf{r}) = u_k exp(j\mathbf{k}\cdot\mathbf{r}) \tag{3.3}$$

The wave vector, $\mathbf{k} = \frac{\mathbf{p}}{\hbar} = 0$, where $\mathbf{p}$ is momentum, $\hbar$ is the reduced Planck constant and u_k the periodic Bloch function resulting from the lattice periodicity.

In three-dimensional semiconductors, the wave vector energy depends on the direction of the wave vector in the lattice, and energy extrema occur with varying $\mathbf{k}$ along the crystal planes. Zinc-blende crystals such as GaAs and InGaAs have three conduction band energy minima: the Γ valley which occurs at $\mathbf{k} = 0$ along the (000) plane; the L valley, which lies along the (111) plane; and the X valley, near the (100) plane. There are additionally three valence band maxima, all at $\mathbf{k} = 0$.

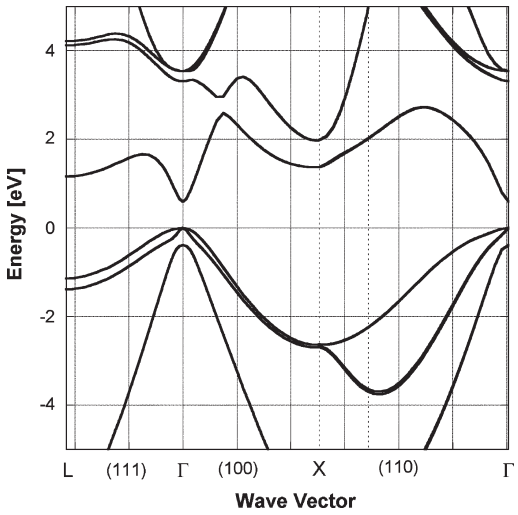


Figure 3.2: Band structure of $In_{0.75}Ga_{0.25}As$. After Ayubi-Moak et al. [22].

Most III-Vs such as GaAs and InGaAs are direct-bandgap semiconductors, as shown in Figure 3.2, in that the conduction band minimum occurs directly above the valence band maximum for a given $\mathbf{k}$.

In free space, the energy of an electron, E, is related to the wave vector as [23]:

$$E = \frac{\hbar^2 k^2}{2m} \tag{3.4}$$

Where m is electron mass. In semiconductors, a similar case applies near the conduction band minimum, where m_e is the effective electron mass [24]:

$$E(k) = E_c + \frac{\hbar^2 k^2}{2m_e} \tag{3.5}$$

As a result, the electron energy is strongly dependent on the value of the wave vector and upon the effective mass. In higher conduction band energy valleys, where effective mass is greater than that in the Γ valley, energy, and so velocity, can therefore be significantly reduced.

Free electrons in semiconductors are subject to both drift and diffusion. Diffusion processes result from random motion due to thermal energy acquired as a result of the environment, and act to drive carriers from densely populated regions to those of low density. This diffusion process results in a net transfer of electrons from one region to another and hence generates a diffusion current.

Drift processes, conversely, result from the application of an external electric field; in transistors, normally due to the application of voltages between the terminals of the device. Electrons, as charged particles, are strongly accelerated by these electric fields, and electron drift results in addition to the thermal processes at work.

Considering an electron moving under the influence of an electric field in a semiconductor, by equating Newton's second law of motion with the applied force from the electric field:

$$m_e \frac{dv}{dt} = q\xi \tag{3.6}$$

Where m_e is electron effective mass, v is the drift velocity resulting from the applied electric field, ξ.

Scattering - a simple overview

In reality, electrons do not propagate ideally in an electric field, and are impeded by collisions and other scattering events which change their energy or momentum. Scattering can have many causes, and generally many processes are simultaneously at work. The effect of scattering is to perturb the electron wave function on each event.

Most scattering is due to aperiodicity or imperfection in the crystal structure, and known generally as defect scattering. Generally, these scattering events are inelastic and electron energy is not usually conserved [25]. Scattering is possible due to unintentional imperfections and dislocations in the semiconductor crystal, and is hence strongly dependent on the growth conditions, strain and substrate. Under well-controlled epitaxial growth conditions, crystal defect scattering is not usually dominant.

Ionised impurity scattering results from coulombic interactions of electrons with the electric fields of dopants which have subsequently become ionised. The situation with ionised impurity scattering is further complicated where carrier density is high by the fact that ionised impurities may attract electrons which act to screen the field, resulting in an electrostatically complex situation [26]. An unfortunate consequence of ionised impurity scattering is that the rate increases with doping.

Neutral impurity scattering occurs due to interactions with unionised impurities, and has a weak contribution.

Ternary semiconductors are also subject to alloy scattering [27] since they are intrinsically not fully ordered. The distribution of the ternary alloy components in the lattice is effectively random, resulting in natural aperiodicity which may scatter electrons. Alloy scattering, however, tends to be a weak effect at room temperature.

Interface scattering [28] is one further defect-based mechanism. Practical interfaces are non-ideal, and subject to both imperfections of roughness and the presence of surface states. Interfacial scattering is usually unimportant for bulk epitaxial growth, but important when considering thin films, conduction near surfaces and metal-semiconductor or insulator-semiconductor interfaces.

At finite temperatures, the crystal lattice also vibrates, with the emission of either low-frequency acoustic phonons or high-frequency optical phonons [25] which may scatter electrons. Optical phonon scattering in particular can have a serious impact on transport, since optical phonons generally have high momentum. Phonon scattering particularly can also lead to intra- or intervalley transfers due to the frequent changes in momentum, resulting in changes in electron effective mass with a correspondingly direct impact on carrier velocities. Optical phonon processes are particularly complex in compound semiconductors due to the polarised nature of the atoms [25].

Electron-electron scattering can also occur at high electron densities, though it is gener-

ally an elastic process. Carrier scattering, however, may cause further interactions with phonon scattering processes.

In general terms, therefore, scattering acts to perturb electron transport, to the detriment of average velocity. Different mechanisms may be dominant under different conditions of field and temperature, and although it may be minimised, scattering is inescapable.

Subtracting a proportional term from Equation 3.6 to account for scattering [29] results in Equation 3.7:

$$m_e \frac{dv}{dt} = q\xi - m_e \frac{v}{\tau_{np}} \tag{3.7}$$

Where τ_{np} is the momentum relaxation time, which describes the mean time between scattering events; generally less than 1 ps. Equivalently, the mean free path describes the distance between successive collision events.

At signal frequencies such that $2\pi f \ll \frac{1}{\tau_{np}}$, $m_e \frac{dv}{dt} \approx 0$. Hence:

$$q\xi \approx m_e \frac{v}{\tau_{np}} \tag{3.8}$$

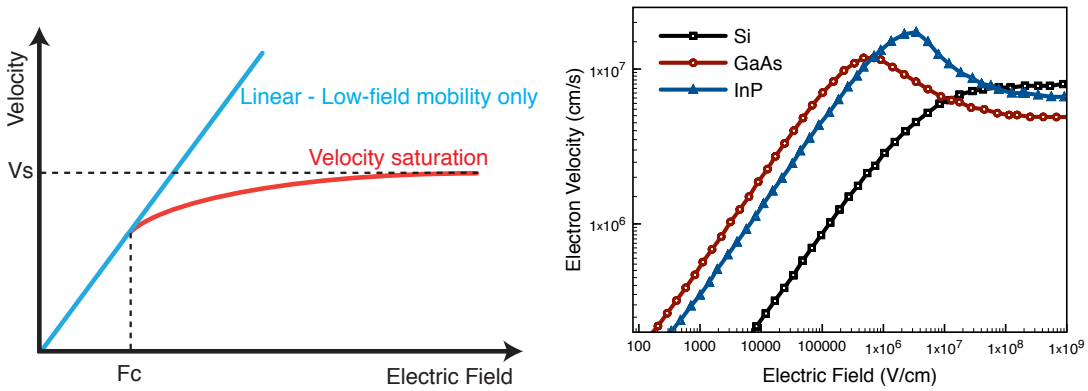
$$\therefore v = \frac{q\tau_{np}\xi}{m_e} = \mu\xi \tag{3.9}$$

Where μ is the low field mobility, which describes the variation of velocity with electric field and is a measure of low-field electron transport.

3.3.2 Field-dependent transport

HEMTs are often operated at high source-drain bias, hence large electric fields. Considering also that the application of a gate voltage can create an electrostatically complex field distribution, it is therefore important to understand the principles at work at various applied electric fields.

At low applied electric fields, electron velocity increases linearly with mobility, according to Equation 3.9. As electron energy increases, however, the scattering rate also increases. As the field increases beyond a critical magnitude, $\mathrm{F_c}$, the electron energy is sufficient to



(a) Schematic showing low-field mobility and velocity saturation effects. Units are arbitrary.

(b) Velocity versus electric field for silicon, gallium arsenide and indium phosphide. Adapted from [30].

Figure 3.3: Field-dependent transport effects in bulk semiconductors illustrating the results of velocity saturation and intervalley scattering on electron velocities.

excite lattice vibrations, resulting in additional phonon emission and increased scattering rates of the so-called "hot" (high energy with respect to the lattice energy) electrons. As a result, the hot electron loses velocity [31] as a consequence of the combined increased scattering probability, and the average drift velocity saturates, no longer dependent on electric field and independent of mobility, as figuratively illustrated in Figure 3.3(a). Optical phonon scattering has been shown to be the principal cause of electron scattering under high-field conditions [32].

In multi-valley semiconductors such as the InGaAs/InAlAs material system, however, there is a further mechanism to consider. As electric field is increased, the electrons gain sufficient energy relative to the lattice to permit transition from the central Γ valley to the L and X valleys [33]. These high-energy valleys feature relatively higher effective masses, resulting in reduced drift velocities as per Equation 3.9 [26]. As electric fields increase, more carriers transit to the upper valleys and velocity eventually saturates at the value associated with the most highly populated valley. Electrons may then subsequently scatter back to the Γ valley. As a function of electric field, velocity therefore peaks, following the saturation curve for the Γ valley, before dropping and reaching a final saturation value determined by the valley occupation distribution of the ensemble carrier population [31]. As a consequence, negative differential mobility [34] can occur as the electric field is increased - an effect harnessed in some resonant devices [35]. Different semiconductors exhibit a range of valley separations and effective masses, and hence

produce various critical fields, peak and saturation velocities, as illustrated in Figure 3.3(b).

The transport situation is further complicated by non-equilibrium effects resulting from the mesoscopic length scales involved in advanced devices, as will be discussed in depth in Section 3.8.

3.4 Heterostructures

The development of epitaxial growth techniques throughout the last two decades of the twentieth century brought the capacity to grow multiple layers of dissimilar semiconductors, known as heterojunctions or heterostructures, allowing the vertical engineering of electronic devices with atomic-level precision; a capability on which devices such as the HEMT are founded.

Since the lattice constant of a semiconductor crystal is dependent on atomic size, there are restrictions on the layers which may be sequentially grown due to the strain that results from large variations in atomic spacing. This strain can induce lattice imperfections, causing increased carrier scattering and poor surface morphology which may hinder device realisation. Whilst layers of dissimilar lattice constant may be grown, the strain must be managed, and constraints exist on possible thicknesses before defects occur [36].

A given ternary or quaternary materials system will exhibit a wide range of bandgaps and lattice constants, not all of which can be realised epitaxially. The arsenide materials system is one of the most versatile, as is demonstrated in Figure 3.4, in which the lattice constants of InGaAs, InAlAs and AlGaAs have been calculated by Vegard's Law [37] and the bandgaps for which have been calculated using the most recent figures from [38].

As indium is added to GaAs, its lattice constant increases between 5.653 Å and 6.08 Å (InAs), whilst adding indium to AlAs (5.661 Å) also increases the lattice constant of InAlAs towards that of InAs.

It is important to note the lattice constants of the ternaries in Figure 3.4 with respect to those of the GaAs and InP binaries which serve as substrates, particularly in the case of $In_{0.53}Ga_{0.47}As$ and $In_{0.52}Al_{0.48}As$, which are lattice-matched to InP. In this case, though the lattice constants are identical, the difference in bandgaps is 0.71 eV.

It is also interesting to note that a strained layer may yield different properties in ac-

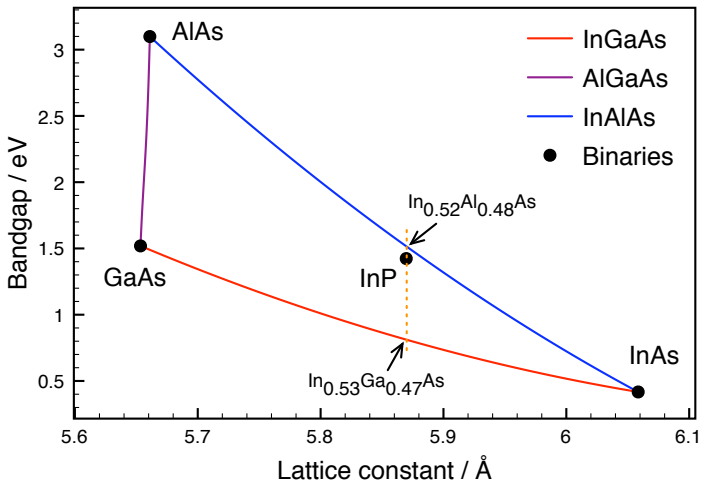


Figure 3.4: Arsenide ternary bandgap and lattice constant variation with composition.

cordance with its altered lattice constant; bandgap and valley separations are modified by the induced strain, as well as the band curvatures [39]. Effective mass variations can result, with corresponding variations in mobility and peak velocities. As a consequence, it has been reported that strained layers incorporated into devices can enhance their performance [40].

The manner in which the buffer layer is grown determines the device type. The simplest case is the lattice-matched device, where the atomic spacing of the channel material is matched to that of the substrate wafer, usually either GaAs or InP, which provide a good lattice match for various ternaries as shown in Figure 3.4. By matching the two lattice constants, lattice dislocations and strain are avoided. Unfortunately, it is usually undesirable to grow lattice-matched channels, which tend to exhibit inferior transport properties.

Metamorphic devices use a graded buffer which gradually varies in lattice constant to accommodate the strain induced by growing a layer of larger lattice constant on a substrate of smaller lattice spacing. Various methods exist, but a popular method uses a graded InAlAs/InGaAs superlattice to grow high-indium channel materials on GaAs.

Pseudomorphic devices utilise a strained channel, usually slightly lattice-mismatched to either the underlying buffer layer. The common use of high-indium channels in pseudomorphic devices results in improved performance due to improved electron transport, but

managing the strain in the mismatched layers is imperative to avoid lattice dislocations. As an example, channels with very high indium content can be grown on buffers lattice-matched to indium phosphide, resulting in improved mobility and saturation velocities. Pseudomorphic approaches can also be adopted on metamorphic buffers, allowing channel materials to be highly mismatched to the substrate by grading the buffer to gradually vary channel strain. The InGaAs/InAlAs materials system will be discussed in greater depth in Section 3.4.

Formation of the InGaAs/InAlAs heterojunction

To consider the mechanisms underlying the formation of heterojunctions, the band structures of the two dissimilar materials must be considered. The Anderson model [41] for band alignment will be used to explain band line-up across the junction, which uses the relative electron affinities of the two semiconductors to calculate the band position. Whilst more physical methods for heterostructure line-up have since been developed [42], the Anderson model is ideologically simpler and more fitting to our purposes, whilst its shortcomings have been reported to be minimal in conceptual terms [43].

We consider first the two semiconductors separately in Figure 3.5(a), where E_c is the lower conduction band edge, E_v is the upper valence band edge and E_f is the Fermi level. The $In_{0.52}Al_{0.48}As$ layer in n-doped, and hence the Fermi level is close to the conduction band edge. In contrast, the Fermi level is near the middle of the bandgap in the undoped $In_{0.75}Ga_{0.25}As$. In addition to their different bandgaps (E_g), $In_{0.52}Al_{0.48}As$ and $In_{0.75}Ga_{0.25}As$ also exhibit different electron affinities (χ), which are defined as the energy required to move a free electron from the conduction band to the energy level of the vacuum (defined as the energy of an electron at rest, far from the solid).

As a result, there exist discontinuities in the conduction and valence band energies between the two semiconductors, equivalent in total to the difference in their bandgaps.

In non-degenerate semiconductors, the conduction band discontinuity is independent of doping, and is specified by the Harrison model to be:

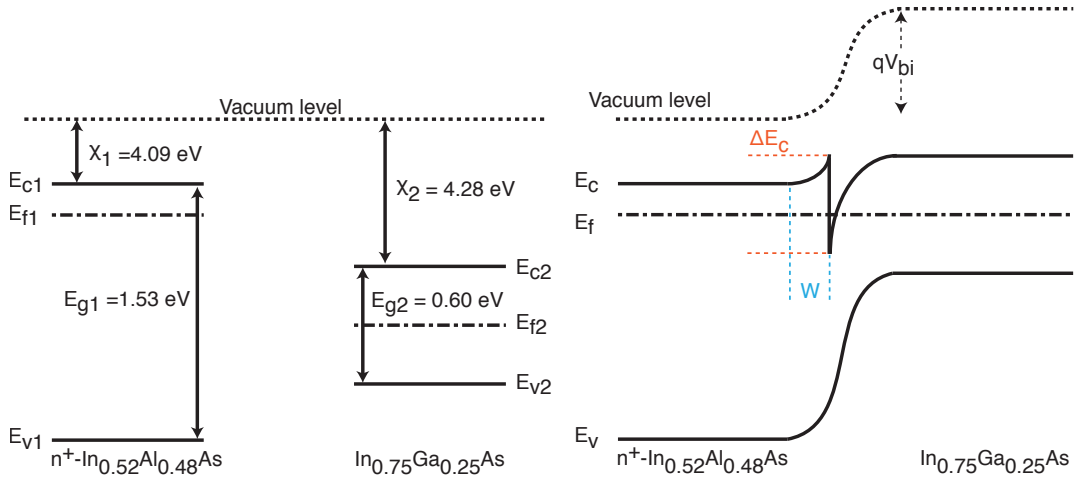
$$E_c = \chi_1 - \chi_2 \tag{3.10}$$

Tersoff, et al. [44], however, showed the offsets to be more accurately determined by the Schottky barrier heights of the two materials, not their electron affinities. Regardless of

the exact magnitude, the relation of the offsets to the bandgaps is:

$$\Delta E_g = E_{g1} - E_{g2} = \Delta \mathrm{E_c} + \Delta E_v \tag{3.11}$$

In isolation, due to their different bandgaps and the heavy doping of the n^+-$In_{0.52}Al_{0.48}As$, the Fermi levels of the two semiconductors are very different. On heterojunction formation under equilibrium conditions, the Fermi level must remain constant across the interface, whilst the vacuum level must remain parallel to the conduction band edge, since the bulk material properties remain unchanged.



(a) Band diagrams for isolated $In_{0.52}Al_{0.48}As$ and $In_{0.75}Ga_{0.25}As$.

(b) n^+-$In_{0.52}Al_{0.48}As$ / $In_{0.75}Ga_{0.25}As$ heterojunction.

Figure 3.5: Formation of the n^+-$In_{0.52}Al_{0.48}As$ / $In_{0.75}Ga_{0.25}As$ heterojunction, as described by the Anderson model.

As a consequence, higher-energy electrons diffuse from the highly-doped $In_{0.52}Al_{0.48}As$ into the undoped $In_{0.75}Ga_{0.25}As$, depleting the $In_{0.52}Al_{0.48}As$ over a region of thickness W and leaving their ionised donors; a net positive space charge. A net negative charge accumulates from the electrons gathering in the InGaAs, eventually reaching equilibrium, when the electron transfer stops.

This dipole sets up an electric field across the heterojunction governed by the built-in voltage, V_{bi}, causing band-bending to occur in the conduction and valence bands of the semiconductors (Figure 3.5(b)). This band-bending, combined with the magnitude of

ΔE_c, results in the formation of a quasi-triangular potential well in the conduction band [45], and is described by the Poisson equation and Gauss's Law [21].

3.4.1 Modulation doping and 2-Dimensional Electron Gas

Since the dimensions of the well formed are similar to the wavelength of the electrons, quantum effects begin to dominate, resulting in the formation of discrete quantised energy levels as a result of the boundary conditions placed on the electron wave function at the edges of the well. Electrons, as a result, sequentially fill the unoccupied minimum energy states. Due to the non-uniform well width as a result of the gradual bending of the conduction band, the energy states become closer together as the well widens, as illustrated in Figure 3.6.

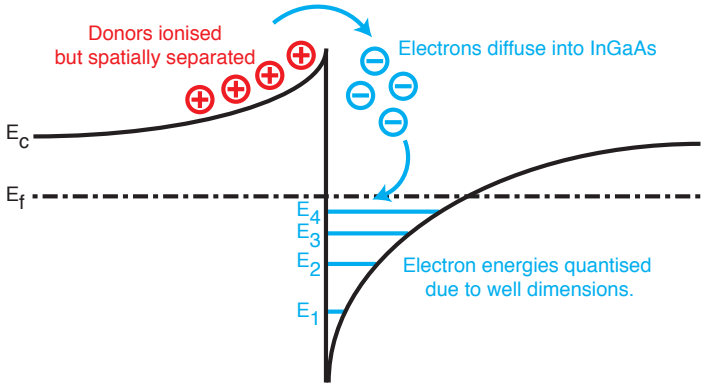


Figure 3.6: Formation of the InGaAs 2DEG.

Since electrons within the well are constrained by this quantum confinement perpendicular to the interface, they are at liberty to move in only two dimensions. This sheet of electrons is known as a two-dimensional electron gas (2DEG). Since the electron accumulation layer in the undoped $In_{0.75}Ga_{0.25}As$ is supplied by the heavily doped $In_{0.52}Al_{0.48}As$, electrons in the 2DEG are spatially separated from their donor impurities and mobility is greatly increased due to the reduction of ionised impurity scattering. This scheme is known as modulation doping, first proposed by Dingle, et al [2]. Modulation doping enables simultaneous high mobility and high electron density; impossible with regular doping. It is noteable that modern HEMTs generally also use an undoped spacer layer, as evident from Figure 3.1, to further separate the channel from the ionised donors, reducing the contribution of remote impurity scattering. Though the use of a spacer

increases mobility as a consequence, it necessitates that a channel doped from above is buried by the combined spacer and barrier thickness, with according scaling issues as detailed in Section 3.8.

By forming a 2DEG layer at the heterostructure interface, a high-mobility channel can therefore be created for a field-effect device, with enhanced mobility only in the directions intended for electron flow, between source and drain.

In order to determine the exact band structure and charge distributions, and therefore the 2DEG population, the Poisson and Schrödinger equations must be solved self-consistently perpendicular to the heterojunction; generally done in software using iterative finite element methods [46] due to the problem's complexity.

The Poisson equation is [47] :

$$\nabla(\kappa(z)\ \nabla\phi(z)) = -q[N_d(z) - N_a(z) + p(z) - n(z)] \tag{3.12}$$

The Schrödinger equation is [47] :

$$\left[-\frac{\hbar^2}{2m_e}\delta + E_c(z) + V_{xc}(z)\right]\xi_i(z) = \epsilon_i\xi_i(z) \tag{3.13}$$

Where $\kappa(z)$ is the dielectric constant at a given position, z, across the heterointerface, $\phi(z)$ is electrostatic potential, $N_a(z)$ is acceptor density, $p(z)$ is hole concentration, ϵ_i is the eigenenergy of the ith sub-band (solution to the wave equation $\xi_i(z)$ at a given position), V_{xc} accounts for electron-electron interactions. All other symbols have their former meanings.

The electron concentration of the 2DEG from a self-consistent Poisson-Schrödinger solution for a generic heterojunction has been determined to be [47] :

$$n(z) = \sum_i \frac{m_e k_B T}{\pi\hbar^2} \ln\left[1 + \exp\frac{E_f - \epsilon_i}{k_B T}\right] |\xi_i(z)| \tag{3.14}$$

The modulation of the 2DEG density by application of a gate bias will be described in Section 3.6.

3.5 Semiconductor interfaces and surfaces.

Understanding HEMT operation requires more than an understanding of transport in semiconductors; devices rely on interfaces to the semiconductor via the various contacts that allow the current flow to be controlled. As such, the interfaces to these contacts are very important.

Interfaces to semiconductors create an interesting situation; transport in a bulk crystal is governed by the carrier wave function as a result of the periodicity of the crystal lattice described by the Kronig-Penney model, but termination of the periodicity significantly affects the band energies. Tamm [48] was the first to realise this possibility, and in doing so, began investigations into one of the least clearly understood solid-state phenomena. He found that it was possible for energy levels to exist whose electron wave functions were restricted to the surface layer. The phenomenon was later investigated by Shockley [49], who discovered that surface energy levels may exist for very small lattice constants at "forbidden" energies outwith the usual allowed energy levels, forming surface bands, which are normally partially filled in semiconductors. He also noted that most metals, also crystalline, would exhibit partial filling of surface states as well.

Bardeen later furthered this work, [50], with a particular focus on the importance of surface states for forming metallic contacts to semiconductors. He noted that, further to Shockley's interpretation, one surface state will be formed for each surface atom, and that they may arise from many more sources than the interruption of the periodic potential; imperfections at the surface, contaminant foreign atoms and impurities. He deduced that the number of surface states is therefore likely to be very large with respect to the number of surface atoms.

The presence of these surface states therefore leads to the creation of a surface charge as the levels become partly filled, as is usual for semiconductors. This charge then acts to repel free electrons, creating a double layer of charge whereby the surface region of the semiconductor becomes depleted of free electrons. Surface states are thoroughly relevant to device design, since they affect not only metal-semiconductor interfaces, but can interact with the surrounding atmosphere to affect reliability, induce extra capacitances, or react in frequency-dependent trapping mechanisms which may affect performance in unpredictable ways.

The Bardeen model was further refined by Heine [51], but the core model remains the

same, and so it is this on which the following discussion centres.

3.5.1 Metal-semiconductor interfaces

The first analyses of metal-semiconductor interfaces were developed by Schottky [52], and do not include any interpretation of surface states, but are a useful starting point. The Schottky model was so successful that rectifying contacts are still usually named Schottky contacts.

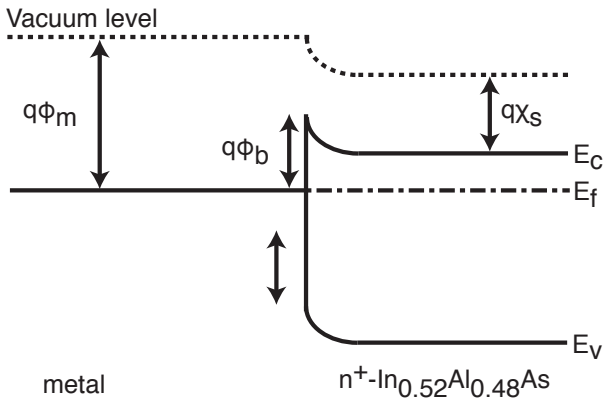


Figure 3.7: Band diagram of an ideal Schottky contact.

As with the formation of a heterojunction, when a metal and n-doped semiconductor are brought into intimate contact, their Fermi levels align, and electrons diffuse from the semiconductor into the metal (where the Fermi level is generally in the conduction band). The diffusion creates an electric field, causing band-bending, until the Fermi levels align under equilibrium conditions.

The band-bending therefore creates a conduction band discontinuity between the conduction bands of the metal and semiconductor, usually referred to as an energy barrier or the Schottky barrier, and illustrated in Figure 3.7.

The height of the ideal Schottky barrier, ϕ_b, is then given by [21]:

$$q\phi_b = q(\phi_m - \chi_s) \tag{3.15}$$

According to the Schottky model, therefore, the magnitude of the barrier should be dependent on the metal work function, ϕ_m and the electron affinity of the semiconductor,

χ_s. For an ideal Schottky contact, therefore, it would be possible to tailor the barrier by controlling both the metal used and the dopant density of the semiconductor.

Bardeen's model [50] of the junction takes surface states into account. Since III-V materials have a large surface state density, it is this model which is more relevant to the understanding of contact formation. In such systems, the interface states are sufficient to pin the Fermi level below the conduction band, with the consequence that the energy barrier becomes independent of the metal work function.

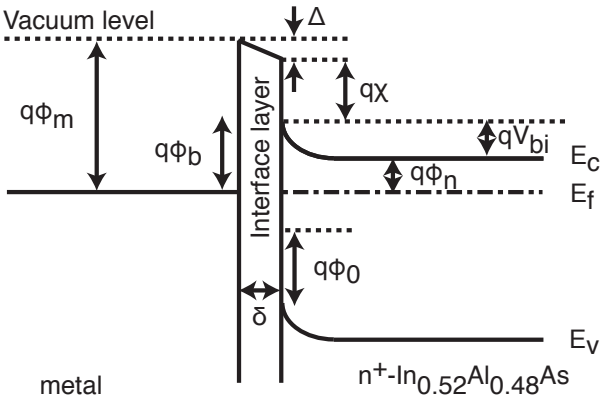


Figure 3.8: Band diagram of a Bardeen-type contact, where surface states pin the Fermi level and the the energy barrier is independent of work function. Adapted from Cowley and Sze [53].

Figure 3.8 shows the pinned-Fermi level case. The region shown between the metal and semiconductor represents the interfacial region of thickness δ defined by the surface state density supporting a potential difference of Δ. ϕ_0 describes the required energy at which surface states must be filled to ensure charge neutrality, whilst ϕ_n is the offset of the Fermi level from the conduction band edge.

If the density of surface states is sufficiently large that $q(\phi_m - \chi)$ can be entirely compensated by the surface states, no diffusion occurs from the semiconductor to the metal, and the process is independent of both work function and doping. As demonstrated by both Figure 3.8 and Equation 3.16, the barrier rather becomes a function of the bandgap and surface state density.

$$\phi_b = E_g - \phi_0 - \phi_n \tag{3.16}$$

As a consequence, the Fermi level in equilibrium is "pinned" to an energy below the conduction band. In the case of lattice-matched InAlAs, this energy is around 0.65 eV [54].

The width of the depletion region formed by the barrier can be calculated using the Poisson equation (Equation 3.17).

$$w = \sqrt{\frac{2\epsilon_s}{qN_d}(q\phi_b - V)} \tag{3.17}$$

Where w is the barrier width, ϵ_s is the dielectric constant of the semiconductor, $q\phi_b$ is the barrier height and V is the applied voltage.

In reality, the barrier is determined by elements of both the Schottky and Bardeen approaches [53]. The dominant contribution to the barrier height is determined by the surface state density and the thickness of the interfacial layer between the metal and semiconductor.

For III-V's, however, a large degree of Fermi pinning is both well accepted and experimentally verified [54, 55]. It is, furthermore, interesting to note that the Schottky effect acts to reduce the effective barrier height close to the metal/semiconductor interface beyond that predicted by a Bardeen-type model, as electric fields due to image charges in the metal cause further band-bending, resulting in rounded energy barriers which decrease in energy approaching the interface. Whilst the effect does occur at equilibrium, it becomes more pronounced under large applied bias [21].

3.5.2 Electron transport across barriers

The energy barriers imposed by metal-semiconductor interfaces must be overcome in order for current to flow. Three main mechanisms exist by which electrons can transit a barrier; thermionic emission, field emission and thermionic field emission; illustrated in Figure 3.9 [21]. In addition, transport may occur by recombination and hole transport via the valence band, but we focus here on majority carrier transport. Whilst in practice, all three of the main mechanisms are likely to play a transport role, the dominant method will be determined by the width of the barrier and its magnitude, which, as discussed in Equations 3.16 and 3.17, are dependent on the intrinsic material properties and the dopant density.

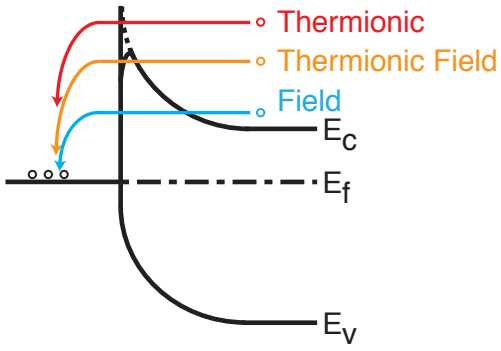


Figure 3.9: Overview of various barrier-transition mechanisms under forward-bias conditions, incorporating the Schottky effect.

Thermionic emission

Thermionic emission relies on thermal energy to excite the electrons with sufficient energy to cross the energy barrier. Thermionic emission generally occurs in low-doped semiconductors, where the energy barrier is too wide for tunnelling to occur, as described by Equation 3.17. As a consequence, electrons must have thermal energy $\mathrm{E} \geq \mathrm{q}\phi_\mathrm{b}$ to transit the barrier.

$$J_{TE} \propto \exp\left(-\frac{E_{barr}}{k_b T}\right) \tag{3.18}$$

The current density under thermionic emission conditions is shown in Equation 3.18; an exponential dependence on the magnitude of the energy barrier, E_{barr}, which takes the barrier values annotated in Figure 3.8 under varying bias, and on temperature, T. In thermal equilibrium, $E_{barr} = qV_{bi}$. As a result, increased temperature will enhance thermionic emission, as will reduced barrier magnitude, caused by a small bandgap or suppressed surface state density (Equation 3.16).

Field emission

Field emission is based on the quantum dimensions of the barrier, and is caused by the tail of the electron wave function extending through a very thin barrier. It therefore tends to occur in highly-doped contacts. Field emission occurs when electrons have insufficient energy to cross the barrier thermionically, but are able to tunnel directly through the

energy barrier, allowing a current to flow.

The tunnelling probability is dependent on the barrier energy magnitude and width, and leads to tunnelling current densities as described by Equation 3.19.

$$J_{FE} \propto \exp\left(-\frac{E_{barr}}{E_{00}}\right) \tag{3.19}$$

Where E_{00} is the tunnelling parameter:

$$E_{00} \equiv \frac{q\hbar}{2}\sqrt{\frac{N_d}{\epsilon_s m_e}} \tag{3.20}$$

Consequently, the tunnelling current is strongly dependent on $\sqrt{N_d}$. Since the barrier width varies with dopant density, the tunnelling probability increases as the barrier thins.

Thermionic field emission

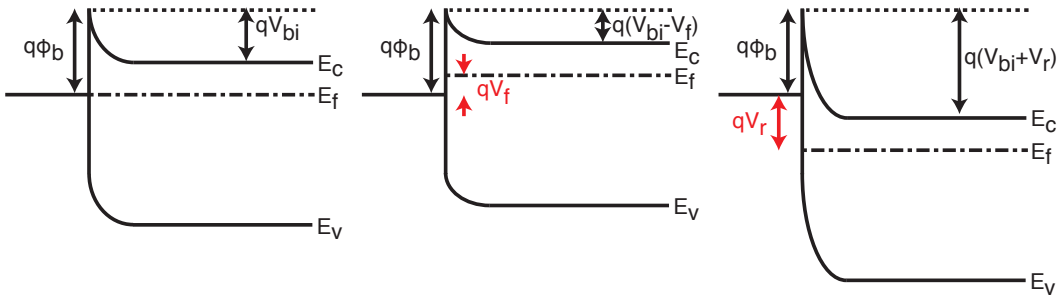
Thermionic field emission usually occurs between the doping densities where thermionic and field emission dominate. It is a combination of the thermal and tunnelling transport phenomena, whereby electrons have insufficient energy to cross the barrier thermally, and the barrier is too wide for direct tunnelling. Electrons with a degree of thermal excitation, however, may may be able to tunnel through the barrier, since its width decreases with increasing energy. This results in the flow of a thermionic field current, varying as [56]:

$$J_{TFE} \propto \exp\left(\frac{E_{barr}}{E_{00}\coth(\frac{E_{00}}{k_B T})}\right) \tag{3.21}$$

Where E_{00} is as expressed in Equation 3.20. Thermionic field emission is therefore expected to increase with increasing temperature, and is strongly dependent on both barrier magnitude and dopant density. In particular, as for field emission, thermionic field emission will increase for a higher dopant density.

3.5.3 Rectifying (Schottky) contacts

In rectifying contacts, the main transport process is by thermionic emission [21], and where a reverse bias is applied, the magnitude of the energy barrier is much greater than the electronic thermal energy.



(a) Zero bias. (b) Forward bias. (c) Reverse bias.

Figure 3.10: Band diagrams of metal/low n-doped semiconductor junction under various bias conditions, where thermionic emission is likely to dominate. After Sze [21].

If the junction is forward-biased (Figure 3.10(b)), the barrier height impeding electron flow from semiconductor to metal is reduced to $q(V_{bi} - V_f)$ as the Fermi level aligns to an energy equal to the applied voltage above the work function of the metal, increasing the probability of electrons acquiring enough energy to cross the barrier into the metal.

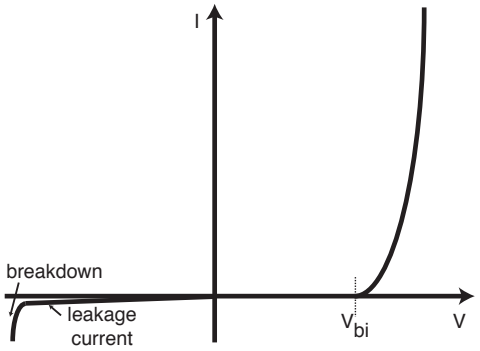


Figure 3.11: Current-Voltage characteristics for a rectifying contact. The transfer characteristics are highly non-linear, and incorporate thermionic emission under forward bias, whilst field emission dominates under reverse bias.

In reverse bias (Figure 3.10(c)), the Fermi level aligns below the metal's work function, increasing the effective barrier magnitude and width, and reducing the probability of electrons transiting into the metal from the semiconductor. Under reverse-bias conditions, the barrier magnitude impeding metal-electron current flow is $q\phi_b$, the barrier due to the surface state density. If the barrier magnitude is small, then carriers are likely to cross by thermionic emission. In the case of a large barrier height, as for $In_{0.52}Al_{0.48}As$, as previously discussed, few electrons will transit, resulting in minimal current flow: a rectifying contact. As the bias voltage made increasingly negative, an increased number of electrons transit by field emission, leading to a relatively small leakage current flow.

Figure 3.11 shows the I-V characteristics of a rectifying contact, where diode-like non-linear behaviour is exhibited. Under forward bias, thermionic emission leads to large currents at applied voltages greater than the built-in voltage, V_{bi}. Under a small reverse bias, field emission leads to small leakage currents, whilst under a very large bias, field emission dominates and a large current flows, referred to as breakdown.

In the case of the HEMT, the gate contact is rectifying in nature, and is deposited directly onto the $In_{0.52}Al_{0.48}As$ barrier layer. As increasingly negative gate voltages pinch off the channel, (Section 3.4.1) one expects the gate leakage, resulting from field emission through the pinned Schottky barrier, to increase as the channel is pinched off. Conversely, a positive bias should reduce the effective barrier height, increasing channel population, but at the risk of establishing a current from gate to source by thermionic emission.

3.5.4 Ohmic contacts

Ohmic contacts, in contrast, exhibit highly linear transfer characteristics, leading to their obedience of Ohm's law. Since thermionic emission results in such highly asymmetric and non-linear transfer characteristics, ohmic contacts must rely entirely on field emission for carrier transport across the Schottky barrier.

As described in Section 3.5.2, the probability of field emission is greatly increased by the reduction of both the magnitude and width of the barrier. Since the width of the barrier is decreased as doping is increased, and the barrier magnitude is dependent on bandgap and surface state density for a Fermi-pinned semiconductor, the control of these parameters is key in creating ohmic contacts.

In the InGaAs/InAlAs system, ohmic contacts are generally defined using highly doped $In_{0.53}Ga_{0.47}As$ as a cap layer in order to form low-resistance contacts. The high level

of doping reduces the width of the barrier, whilst the use of a cap layer reduces the magnitude of the barrier, due to the narrow bandgap of $In_{0.53}Ga_{0.47}As$ and the pinning of its Schottky barrier at around 0.3 eV below the conduction band edge. The extreme to this approach lies in the use of InAs as a cap layer, since its surface is pinned in the conduction band, eradicating the barrier entirely [57].

Whilst low-resistance contacts have been achieved in HEMTs using these approaches [58, 59], most low-resistance contacts are formed by annealing to produce an alloyed contact, which act to reduce the energy barrier by producing very highly-doped regions formed during the annealing process. Commonly, gold/germanium metallisations are used, with gold as the base layer, where the germanium diffuses into the semiconductor, locally doping it and reducing the barrier energy [60]. Recipes involving nickel have also become popular [61], with the nickel forming a barrier layer which contains the alloyed mixture, reducing out-diffusion, lowering the required annealing temperatures [62]. In general terms, alloyed ohmic contacts become unstable with high temperatures [63], though the upper annealing temperatures for these recipes are usually close to those used during the epitaxial growth of the material to which the contact is made.

Despite the maturity of low-resistance contact formation on III-Vs, there is a constant drive for the continued reduction of contact resistances, which become increasingly significant as devices are scaled [4].

3.6 Device operation

With an understanding of the theory behind electron transport in semiconductors, channel formation and modulation and transport across barriers, it is now possible to apply this knowledge to HEMT operation. Electron density and basic electron flow in the channel have been discussed, but this must yet be applied to appreciate ensemble current flow under varying bias conditions.

We first consider the modulation of the channel electron population, introduced in Section 3.4.1, by a Schottky gate.

As discussed in Section 3.5.3, the gate controls the energy band geometry. It therefore also controls the 2DEG density in the channel. At zero gate bias, the channel is populated with a high-density 2DEG resulting from the channel doping. On the application of a negative gate voltage, the surface Schottky barrier is increased, bending the conduction

band and decreasing the 2DEG density. A positive gate bias will reduce the barrier and so increase channel population.

For some given threshold voltage, V_{th}, the channel becomes completely depleted, and current flow between source and drain is impossible. In this condition, the channel is said to be "pinched off", and the transistor's resistance is extremely high.

The first simple model for charge control in the channel by a Schottky gate contact was developed by Delagebeaudeuf and Linh [64], and described the capacitive coupling of the gate to the channel by approximating the potential well as triangular to allow analytical solutions of the sub-band energies. They assumed that for $0 \leq n_s \leq n_{so}$ (where n_s is 2DEG sheet electron density and n_{so} is its known maximum density):

$$qn_s = \frac{\epsilon}{d}(V_{gs} - V_{th}) = C_s(V_{gs} - V_{th}) \tag{3.22}$$

Where C_s is the 2DEG capacitance per unit area and V_{gs} is applied gate voltage.

The threshold voltage is then defined as the gate voltage at which n_s linearly approaches zero [65] :

$$V_{th} = \phi_b - \frac{qN_d}{2\epsilon}d_n^2 - \Delta E_c - \Delta E_f \tag{3.23}$$

Where ϕ_b is the Schottky barrier potential, d_n the thickness of the doped layer, ΔE_c the conduction band offset in the junction, ΔE_f the offset of the Fermi level to the bottom of the narrow-band conduction band at the interface and other symbols have their former meanings.

This linear model is well-behaved at very low-temperatures, but fails under real operating conditions. In reality, charge control is non-linear, caused by the simultaneous modulation of donors as well as electrons, affecting the electrostatics and rendering the 2DEG capacitance variable. As a consequence, modern models the likes of [46] use a self-consistent approach which model all charge precisely by a full solution of the Poisson-Schrödinger equations.

We now consider the intrinsic HEMT as shown in Figure 3.12, where x specifies the position in the channel from the start of the intrinsic gate control region on the source

side. Areas outwith the intrinsic region are known as access regions, and in addition to the intrinsic region, make up the complete, extrinsic device.

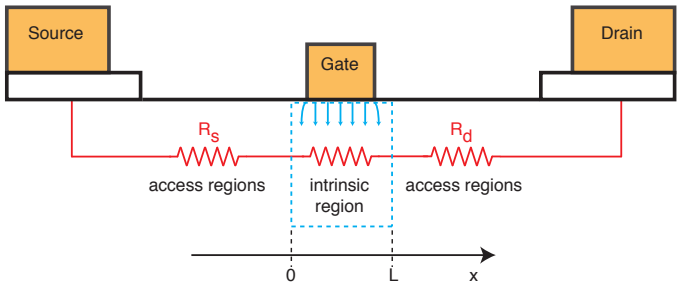


Figure 3.12: Overview of HEMT intrinsic and access regions of the channel. The three areas are treated as separate resistances.

3.6.1 Long-channel electron transport

We consider the long-channel case first, where the electron velocity varies linearly with constant mobility until saturation, as first interpreted by Delagebeaudeuf and Linh [64]. The channel 2DEG is controlled by the application of an electric field, induced by the applied gate voltage. As a bias voltage is applied between source and drain, the electric field in the channel will vary, dependent on the position in the channel, with a corresponding change in the effective gate voltage which acts upon the 2DEG [64].

The effective gate voltage is then:

$$V_{eff}(x) = V_{gs} - V_c(x) \tag{3.24}$$

Substituting Equation 3.24 into Equation 3.22 for the 2DEG electron density, we obtain

$$qn_s(x) = \frac{\epsilon}{d}(V_{gs} - V_c(x) - V_{th}) \tag{3.25}$$

And the intrinsic ensemble current, I_{int} at point x in the channel is therefore:

$$I_{int}(x) = qn_s(x)Wv(x) = \frac{\epsilon}{d}(V_{gs} - V_c(x) - V_{th})Wv(x) \tag{3.26}$$

Where W is the channel width, as previously described in Figure 3.1. The current in the channel is therefore strongly dependent on the device dimensions, the 2DEG capacitance, bias conditions and crucially the electron velocity.

Low-field condition : drift and mobility

Under low-field conditions, as described in Section 3.3.2, electron velocity varies linearly with electric field, according to Equation 3.9. Consequently, we acquire

$$I_{int}(x) = \frac{\epsilon}{d}(V_{gs} - V_c(x) - V_{th})W\mu\xi(x) \tag{3.27}$$

Under low-field conditions, therefore, current should increase with applied electric field, implying increases in current for increased applied bias voltages. Electric field, $\xi = \frac{dV_c}{dx}$, and we approximate the intrinsic region to the gate length, which yields:

$$I_{int}(x) = \mu W\frac{\epsilon}{d}(V_{gs} - V_c(x) - V_{th})\frac{V_c(L_g) - V_c(0)}{L_g} \tag{3.28}$$

Consequently, one expects increased currents in the mobility-controlled regime as the gate length is reduced, with the proviso that electrostatic integrity is maintained such that channel voltages remain independent of gate length.

In considering only the intrinsic region, we have neglected the access or "parasitic" regions of the channel on the source and drain sides of the gate, shown in Figure 3.12, which also play a key role in the definition of channel current. Applying Kirchoff's Voltage Law:

$$V_c(0) = R_s I_{int} \tag{3.29}$$

$$V_c(L) = V_d - R_d I_{int} \tag{3.30}$$

Substituting Equations 3.29 and 3.30 into 3.28, we acquire:

$$I_{int} = \mu W\frac{\epsilon}{dL_g}(V_{gs} - V_c(x) - V_{th})(V_d - R_d I_{int} - R_s I_{int}) \tag{3.31}$$

Rearranging:

$$\frac{V_d}{I_{int}} = R_s + R_d + \frac{L_g d}{\mu W \epsilon (V_{gs} - V_c(x) - V_{th})} \tag{3.32}$$

Consequently, the channel resistance is defined by the only parameter not determined by geometry or materials; the gate voltage.

Velocity saturation

Under velocity saturation conditions, $v = v_s$, and the definition in Equation 3.32 no longer holds.

By integrating Equation 3.27 with respect to x for the channel voltage, and assuming the velocity to be unsaturated on the source side of the gate, but saturated on the drain side, Delagebeaudeuf and Linh find, substituting Equation 3.29, the channel current to be:

$$I_{sat} = \frac{W \epsilon v_s}{d} \left(\sqrt{\xi_c^2 L_g^2 + (V_{gs} - V_{th} - R_s I_{sat})^2} - \xi_c L_g \right) \tag{3.33}$$

Where I_{sat} is the ensemble saturation channel current, independent of channel position. The channel current is dominated by saturated electron velocity and device geometry, but also has a strong dependence on the critical electric field for velocity saturation, ξ_c, and, crucially, the gate length.

Assuming a short gate length device, Equation 3.33 reduces to:

$$I_{sat} = \frac{W \epsilon v_s}{d} ((V_{gs} - V_{th} - R_s I_{sat}) - \xi_c L_g) \tag{3.34}$$

The saturation current is therefore linearly dependent on the gate length. It is therefore expected that as gate length decreases, the saturation current will increase, assuming no major changes in the saturation electric field or threshold voltage.

It is worth noting the importance of the source resistance in this expression, since it defines the voltage drop across the intrinsic region, with a direct impact on electric field, and hence the resultant channel current.

3.6.2 Gate voltage modulation and I-V characteristics

The model considered thus far explains the linear and saturation characteristics of HEMT $I_{ds}/(V_{ds}, V_{gs})$ characteristics. The resulting velocity-saturated current is then modulated by the application of the gate voltage, as described in Section 3.4.1, resulting in I_{ds}/V_{ds} characteristics stacked vertically by the gate voltage. Figure 3.13 shows the idealised characteristics. It should be noted that the effects of output conductance, detailed in Section 3.6.2, are also included.

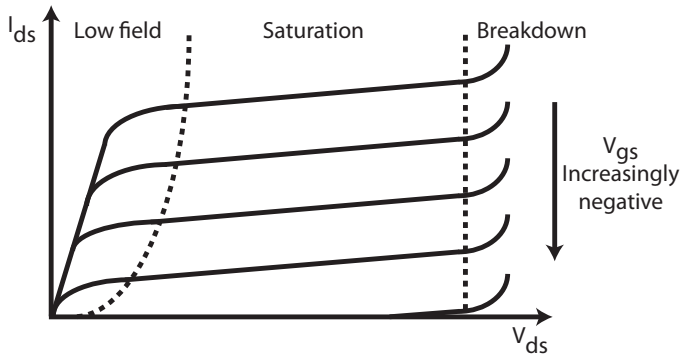


Figure 3.13: Idealised HEMT I-V characteristics, showing linear (low-field), saturation and breakdown regions.

The efficiency with which the application of the gate voltage modulates the channel is a figure of merit of great importance, since it has an effect on the ultimate frequency performance of the device. The most effective measure is transconductance.

Transconductance, $\mathbf{g_m}$

Transconductance is defined as the rate of change of drain current with applied gate voltage:

$$g_m = \frac{\delta I_{ds}}{\delta V_{gs}} \tag{3.35}$$

Using the same analytical long-channel linear charge control approximation as previously and differentiating Equation 3.33, we find the intrinsic transconductance to be [65]:

$$g_{m0} = \frac{\delta I_{ds}}{\delta V_{gs}} = \frac{W\epsilon v_s}{d} \frac{V_{gs} - V_{th} - V_c(0)}{\sqrt{(V_{gs} - V_{th} - V_c(0)) + (\xi_c L_g)^2}} \tag{3.36}$$

$$= WC_s v_s \frac{V_{gs} - V_{th} - V_c(0)}{\sqrt{(V_{gs} - V_{th} - V_c(0)) + (\xi_c L_g)^2}} \tag{3.37}$$

Consequently, the 2DEG-gate capacitance, C_s, defined by the gate-channel separation, directly impacts on transconductance, whilst the key extrinsic component is, again, the source resistance, since it will define $V_c(0)$ in the intrinsic expression.

For completeness, considering the short gate length simplification of Equation 3.34,

$$I_{sat} = \frac{\frac{W\epsilon v_s}{d}}{1 + \frac{R_s W\epsilon v_s}{d}} (V_{gs} - V_{th} - \xi_c L_g) \tag{3.38}$$

$$= \frac{WC_s v_s}{1 + R_s WC_s v_s} (V_{gs} - V_{th} - \xi_c L_g) \tag{3.39}$$

$$\therefore g_m = \frac{WC_s v_s}{1 + R_s WC_s v_s} \tag{3.40}$$

As for the long-channel case, therefore, in short gate length devices, electron velocity, gate-channel separation and source resistance critically define transconductance.

Considering a more intrinsically useful concept for charge modulation, however, the modulation efficiency, η is useful [65]. In a real capped structure, whilst the bulk of the charge is due to the 2DEG population, there are also electrons bound to donors, n_{bound} and free electrons outwith the channel, n_{free}, and therefore moving at comparatively low velocities. We continue to assume the long-channel saturation model. The modulation efficiency specifies the efficiency of modulation of the channel charge with respect to the corresponding change in drain current.

$$\frac{\delta I_{ds}}{\delta Q_{tot}} = \frac{\delta(qv_{sat}n_s)}{\delta(q(n_s + n_{bound} + n_{free}))} \tag{3.41}$$

$$= v_{sat}\frac{\frac{\delta n_s}{\delta V_{gs}}}{\frac{\delta(n_s+n_{bound}+n_{free})}{\delta V_{gs}}} \tag{3.42}$$

$$= v_{sat}\eta \tag{3.43}$$

The modulation efficiency, η, is therefore defined as the ratio of the rates of change with respect to gate voltage of 2DEG charge to that of total charge [65].

$$\eta = \frac{\frac{\delta n_s}{\delta V_{gs}}}{\frac{\delta(n_s+n_{bound}+n_{free})}{\delta V_{gs}}} = \frac{C_s}{C_{total}} \tag{3.44}$$

If $\eta = 1$, the HEMT is operating at 100% modulation efficiency, where only the 2DEG charge is modulated by the gate.

According to the linear charge control model,

$$qn_s = C_s(V_{gs} - V_c(0) - V_{th}) \tag{3.45}$$

Substituting Equation 3.45 into Equation 3.36 renders the transconductance as:

$$g_{m0} = C_s v_{sat}\frac{\frac{qn_s}{C_s}}{\sqrt{\frac{qn_s}{C_s} + (\xi_c L_g)^2}} = C_s v_{sat}\frac{1}{\sqrt{1 + \left(\frac{n_c}{n_s}\right)^2}} \tag{3.46}$$

$$= v_{sat}\frac{qn_s}{\sqrt{\frac{qn_s}{C_s} + (\xi_c L_g)^2}} \tag{3.47}$$

Where $n_c = \frac{\xi_c C_s L_g}{q}$, and is approximate to the change in 2DEG density across the gate at saturation. As a consequence, Equation 3.46 shows the intrinsic transconductance can be understood as directly dependent on the modulation efficiency in addition to the 2DEG capacitance and electron density, whilst Equation 3.47 more clearly reflects that

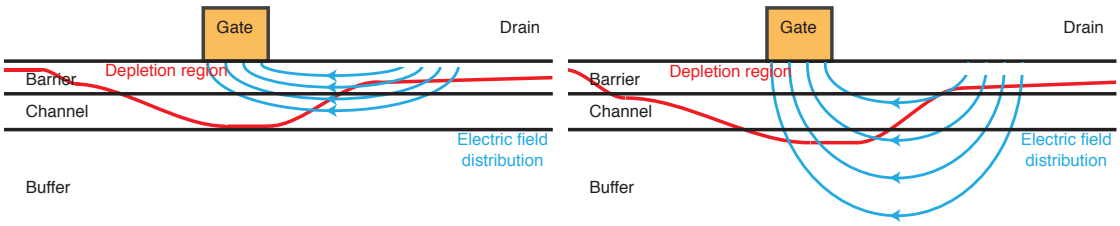
the reduction of the gate length and reduction of the gate-channel separation will increase transconductance.

As a result, the intrinsic transconductance can be improved by increasing the gate's capacitive coupling to the 2DEG, a scaling issue discussed in Section 3.8, by improving the 2DEG density, a materials issue largely defined by the compositional fraction of the channel, the quantum confinement and doping, and by increasing carrier velocity.

In more general terms, increased transconductance leads to a greater separation of the I_{ds}/V_{ds} traces for various fixed gate voltages and saturation current at zero gate bias.

Output Conductance, g_{ds}

In practice, and as reflected in Figure 3.13, the current in the saturation region is not constant with drain-source voltage. Under high drain bias, electric fields increase particularly aggressively at the drain end of the gate. The effect is to reorient the electric field distribution with increased curvature towards the buffer, as schematically presented in Figure 3.14.



(a) Low field. (b) High field.

Figure 3.14: General illustration of electric field distribution around the gate, extending towards the buffer with increasing curvature as drain bias is increased.

As a consequence, real space transfer of hot electrons into the buffer may occur. This results in increased conduction band state occupancy into the buffer, despite the conduction band offset present at the heterointerface; particularly probable at the drain end of the gate. Conduction can then also occur through the buffer, though transport is saturated in the channel. This leads to an effective spreading of the channel, leading to a buffer current which increases with electric field and adds to the saturated channel current [66].

The effect is described by the output conductance, g_{ds}, defined as:

$$g_{ds} = \frac{\delta I_{ds}}{\delta V_{ds}} \tag{3.48}$$

The output conductance therefore defines the gradient of the I_{ds}/V_{ds} curve, and is the inverse of the effective output resistance of the transistor.

Breakdown

In the breakdown regime, the output conductance, usually constant, rapidly increases, and drain current increases rapidly. Breakdown can occur in both the pinched-off and open channel conditions, and is thought to occur as a consequence of both thermionic emission and impact ionisation phenomena, though one effect will dominate in each case.

Off-state breakdown is thought to occur mainly due to thermionic emission directly from the gate as a result of rapid electron heating under low-to-medium drain bias, where electrons penetrate the barrier at the drain end of the gate and flow through the $In_{0.52}Al_{0.48}As$ barrier or channel layer to the drain contact. Under higher bias, the breakdown phenomenon is thought to be two-step [67], and thermionic emission processes may occur with increased electron temperature. Hot electrons are injected into the channel, where they participate in impact ionisation interactions, gradually relaxing energy and causing an avalanche process [68]. Electrons flow towards the drain, holes to the gate or source, and recombine.

In the on-state, the same processes are at work, but impact ionisation always dominates, and likely occurs in the channel [69–71].

InGaAs/InAlAs HEMTs are particularly susceptible to breakdown in both conditions as a consequence of the comparatively narrow bandgap of the channel, barrier and buffer materials, which increases the probability of impact ionisation, and the smaller Schottky barrier at the gate. Various techniques have been investigated to minimise these effects, such as composite channel devices which rely on real space transfer of electrons into a wider-bandgap material to reduce impact ionisation effects, or modulating the channel thickness to control quantum confinement [67, 72]. Some groups have also worked on surface-depleted caps and recessing geometry to more intricately control electric field [73].

Breakdown limits the potential maximum voltage of the transistor, restricting its attainable controlled output currents and hence application.

The Kink Effect

The kink effect in HEMTs is an undesirable non-linearity which introduces a spurious increase in drain current with its onset at given gate and drain bias conditions. The increase is undesirable due to its unpredictability and additionally induces reduced gain and increased noise at high frequencies [74].

The physical basis for the kink has been the subject of some debate. The current understanding [75] involves the phenomena of trapping, including that due to surface states, and impact ionisation effects similar to those at work during breakdown; in effect, the kink is related to the variable charging of surface states and other interface and buffer traps.

The kink is therefore temporally-dependent. Neglecting impact ionisation, channel charge is imaged with an opposing polarity by charged surface states, interface charges and dopant ions. With the onset of impact ionisation, itself a field-dependent effect, holes are generated by the interactions of high-energy electrons at the drain end of the gate. These holes are attracted to the source, accumulating around the source in the channel, resulting in a steady-state channel hole distribution which modifies the potential profile in the vicinity of surface states, increasing electron density and reducing source resistance [76].

To satisfy the surface Fermi pinning (Section 3.5.1), the hole quasi-Fermi level bends, resulting in hole current flow to the surface and into the buffer. The charge at these locations therefore changes, and must be balanced by a corresponding change in the channel electron population, raising the channel potential and resulting in a kink voltage.

The kink effect is therefore most pronounced for wide recesses, where there is a greater total number of surface states to pin the Fermi level; as indeed experimentally evidenced [77, 78]. Many devices now use indium phosphide as a cap to the barrier, which is etched as part of the gate process, in order to reduce the effect of Fermi level pinning of the barrier [12, 79].

It has, however, also been noted that control of field distribution and doping levels also significantly affect the influence of these mechanisms on the presence of the kink.

3.7 Device elements and the small-signal equivalent circuit

The discussion until now has focussed on the theoretical operation and performance of the HEMT in terms of electron transport dynamics in the immediate intrinsic device region - the region of the device below the gate where the gate modulates the channel. In real devices, however, performance is driven by both the intrinsic device and the access regions surrounding it; additional sections of the channel and cap and contact pads, and the secondary effects of their relative geometries. In reality, therefore, these "extrinsic" regions degrade the performance of the intrinsic device by the introduction of additional resistive or reactive effects. At d.c., but particularly at high frequencies, these additional components play a significant role in defining performance.

High frequency transistor operation is frequently understood in terms of equivalent circuit models, which allow the operation of the device to be characterised in terms of lumped circuit elements. An equivalent circuit allows the behaviour of the device to be modelled, and the influence of the various circuit elements to be understood. Additionally, a crucial role of the circuit model is the extraction of the intrinsic device characteristics from those of the extrinsic device, allowing feedback on process parameters to be optimised or device elements which particularly dominate performance. Each of the circuit elements has a physical basis in the device composition, and the circuit is comprised of standard active and passive circuit components.

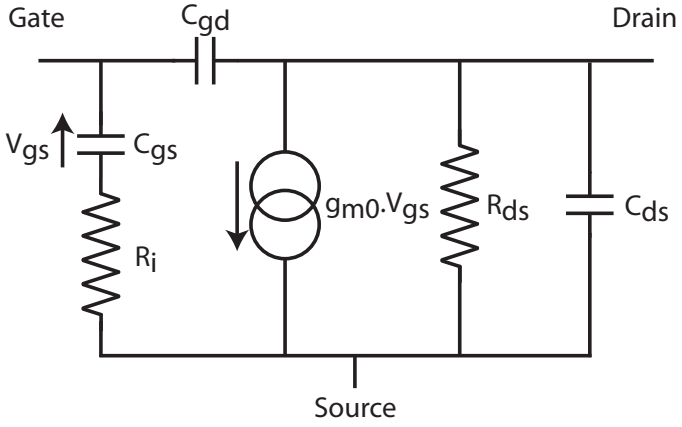


Figure 3.15: Equivalent circuit of the intrinsic device region, showing voltage controlled current source, gate capacitances, intrinsic channel resistance and the combined output impedance. After Ladbrooke [80] & Wolf [81].

The intrinsic device model is centred around a voltage-controlled current source in parallel

with a resistance, where the current source represents the modulated source-drain current flowing in the channel and the resistor represents the effective output resistance, the inverse of the output conductance. This combination models the modulation of the channel current by the gate. The current generated by the source is defined as the product of the intrinsic transconductance, g_{m0}, mentioned in Section 3.6, and the voltage across part of the gate capacitance, also in parallel.

The two capacitances which comprise the remains of the intrinsic model represent the capacitive coupling of the gate to the channel. Since there is significant electrostatic contrast between the source and drain sections of the channel at either side of the gate, splitting the gate capacitance into two allows more representative modelling. The precise variations will depend on the exact depletion region geometries. The resistance of the intrinsic section of the channel is further modelled by a resistance in series with the source-end gate capacitance, whilst a capacitance also arises between the source and drain ends of the intrinsic region, as a consequence of their different electron densities.

As a consequence, the intrinsic device region is modelled by six elements, listed in Table 3.1 and detailed in the circuit diagrams of Figure 3.15 [80].

Circuit element	Symbol	Description
Current source	$V_{gs}.g_{m0}$	Models channel current modulation, transconductance. Controlled by voltage across gate capacitance.
Output resistance	R_{ds}	Inverse of output conductance.
Source-end gate capacitance	C_{gs}	Controls current source, half of distributed gate capacitance.
Drain-end gate capacitance	C_{gd}	Other half of distributed gate capacitance.
Intrisic channel resistance	R_i	Models the finite conductance of the channel.
Source-drain capacitance	C_{ds}	Models transverse capacitance along channel due to varying electron density.

Table 3.1: Intrinsic equivalent circuit components.

In addition, there will be a transconductance delay time associated with the current source, which also has physical significance. Electrons crossing the intrinsic region ideally contribute directly to the drain current, but in the case where the gate voltage changes and the potential difference between gate and channel varies, the depletion width under the gate will also vary, with the effect of these electrons also charging the gate capacitance across the depletion region, giving rise to charging delays. The propagation time along

the gate width will also play a factor. This is modelled as a "lag" of the current source response to its input gate voltage [80].

In addition to the components which make up the intrinsic model, the extrinsic access regions give rise to several further passive components, referred to as "parasitics" since they parasitically degrade optimum device performance.

3.7.1 Parasitic elements and effects on performance

Contributions outwith the intrinsic region originate primarily from the contacts and adjacent sections of the channel. Whilst the optimisation of the intrinsic region will have a critical effect on transistor performance, if the parasitic components are neglected, the performance will be significantly impaired.

It is worth noting that each component of the model is a lumped circuit element, representing the distributed transmission line effects present across the width of a device. At RF frequencies, the gate and drain contacts in particular act as transmission lines rather than as perfect conductors, and hence are described in the model by a lumped element approach. The lumped element method thereby aims to approximate the phase delays and attenuations of these transmission lines, which may not also be perfectly impedance-matched to the measurement system, resulting in further loss.

Each of the contacts will predominantly add a parasitic resistance and inductance associated with the distributed transmission line effects, dependent on the physical device geometry, width and layout, as well as the physical geometry of each of the contacts. In particular, as the metal stripes becomes narrower, their resistance greatly increases. As a consequence, there will be an inductance and resistance in series at the contact points of the intrinsic circuit. There will also be capacitances arising between all the contacts, and a parasitic capacitance is therefore added from the gate and drain to the source, and between gate and drain. These parasitics include the effect of the contacts near the intrinsic region, from the larger pad regions and the feeds connecting the two. The complete small signal equivalent circuit is shown in Figure 3.16.

Although the simplest device layout is the architecture shown in Figure 3.1, in reality, multiple-finger layouts are generally used to optimise the performance at high frequencies, and as a result, a completed device may not immediately resemble this simple structure. By using multiple gate fingers, the effective high frequency gate resistance is considerably reduced for a given gate width [80]. It is additionally worth noting that high frequency

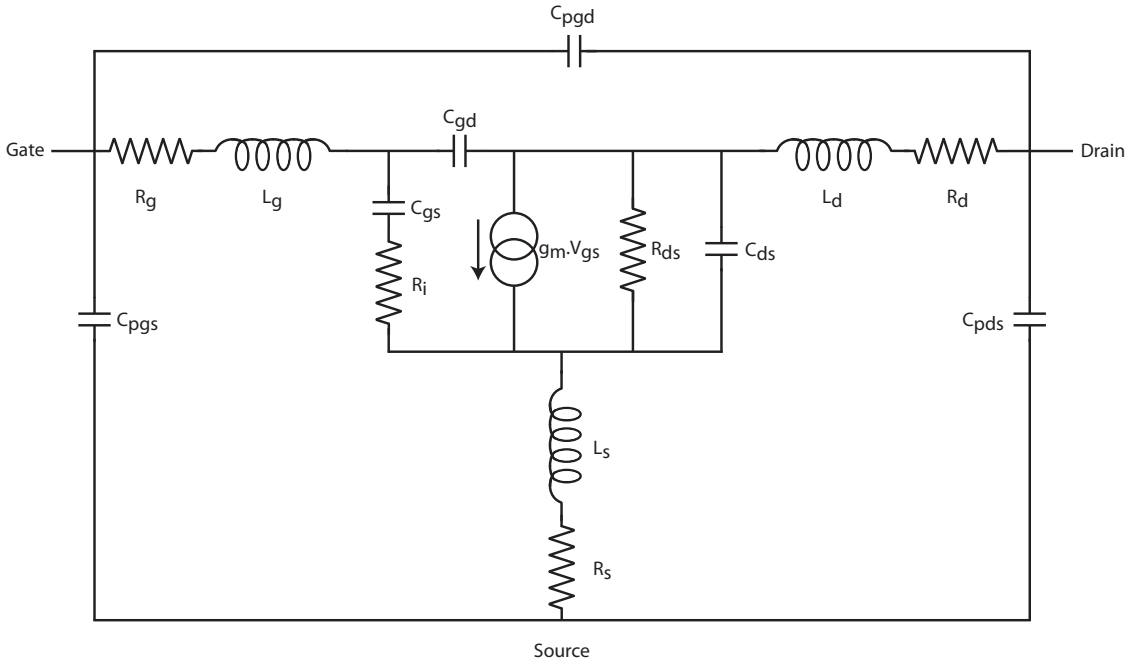


Figure 3.16: Complete extrinsic equivalent circuit, including the parasitic resistances, inductances and capacitances arising from the contacts and additional channel resistance.

gate resistance is approximately one third that at d.c. [81].

The source and drain resistances, in particular, however, are significantly more complex. The source and drain resistances must include more than just the effects from the transmission line characteristics of the contacts, since the entire region from signal input to the intrinsic region is included in the parasitic resistances. Additional resistances associated with the ohmic contacts must therefore also be included.

Parasitic resistances

Each of the ohmic contacts has an associated resistance which will be dominated not by the resistance of the metal contact, but by the lossy path of current flow to the intrinsic region of the channel. This is effectively a resistive chain and is outlined in Figure 3.17.

A parasitic resistance is effectively comprised of the contact resistance, defined by the quality of the ohmic contact at the Schottky barrier at the metal/semiconductor interface, as previously described in Section 3.5.4, and two resistances arising from the sheet

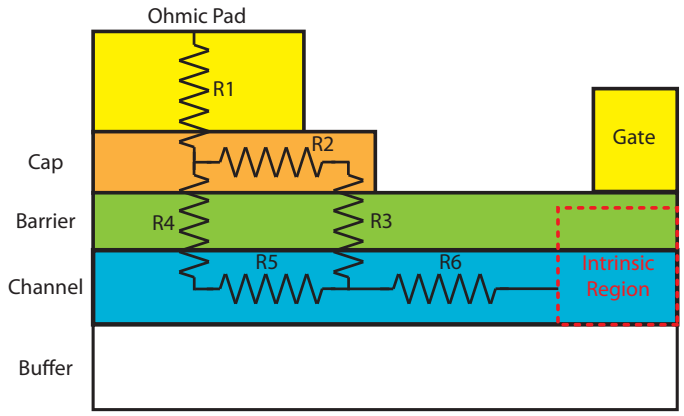


Figure 3.17: Parasitic resistances and their origins in contact and relative sheet resistances of the cap and channel.

resistance of the cap and channel. Considering the path of current flow in this system, a degree of parallel conduction will occur between the cap and channel, the ratio of which will be defined by the relative resistances of the two, and by the linearity of the ohmic contact resistance between the cap and channel at different depths.

The total parasitic resistance of the system will therefore be defined as:

$$R_p = R_1 + \frac{(R_2 + R_3)(R_4 + R_5)}{R_2 + R_3 + R_4 + R_5} + R_6 \tag{3.49}$$

The magnitude of resistances R_3 and R_4 are determined by the magnitude of the conduction band barriers formed between the cap, barrier and channel layers as a consequence of Fermi level alignment of their heterostructures. These barriers impede electron flow between layers and create effective resistance. Their magnitude, and hence the total magnitude of the contact resistance, will be dependent on the design of the layer structure.

Most annealed ohmic contacts are very low resistance as a result of the annealing process, which causes local doping and alloying of the semiconductor under the pad. The effect is to greatly reduce the magnitudes of R_1 and R_4 in Figure 3.17. In regions outwith the area of the contact, however, the resistance R_3 is generally much larger. As a consequence, most current is likely to flow $R_1 \rightarrow R_4 \rightarrow R_5 \rightarrow R_6$ as a result of this reduction, and this path will most significantly affect the total contact resistance. It is to be noted that this generalisation is strongly dependent on the relative resistances of the cap and channel, and therefore the relative magnitudes of (R_2+R_3) and (R_4+R_5).

In the case of non-annealed contacts, however, there is no deliberate alloying of the ohmic metal, and so $R_3 = R_4$. As a result, the path of current flow will be defined by the ratio of R_2 and R_5, and conduction may be genuinely parallel in the entire parasitic region under the pad to the end of the cap. The relative sheet resistances of the cap and channel will therefore specify the contact resistance in addition to surface rectification effects.

The magnitude of conduction band barriers can be tailored by the variation of the compositional fraction of materials, or of doping levels in the respective layers.

It is also worth considering that the electron flow characteristics in the cap will affect, to some extent, the resulting gate-source and gate-drain parasitic capacitances. If the cap is heavily populated to the ends of the exposed cap region outwith the contact, the result is similar to an effective extension of the pad. If this results in a decreased offset between the ohmic region and the gate, parasitic gate capacitances, C_{pgs} and C_{pgd}, as in Figure 3.16, will increase.

Performance effects

The parasitic components have myriad effects on performance, particularly at high frequency, where the reactive contributions become increasingly significant, as will be discussed in detail in Section 3.7.2. At d.c., however, the parasitic resistances play a key role in defining the transfer characteristics of the device.

We consider a HEMT in a conventional common-source circuit application, as in Figure 3.18 [82].

It is clear that the effect of the additional parasitic resistances is to incur a voltage drop from the applied source-drain bias. Considering the HEMT's output characteristics in the linear regime (Figure 3.13), the larger the parasitic resistances, the greater the voltage that must be applied to induce a given drain current. This will manifest as a decrease in the gradient of the I_{ds}-V_{ds} curve in the linear region and hence decreased output conductance.

The intrinsic transconductance was defined in Section 3.6.2, Equation 3.35 as:

$$g_{m0} = \frac{\delta I_{ds}}{\delta V_{gs}} \tag{3.50}$$

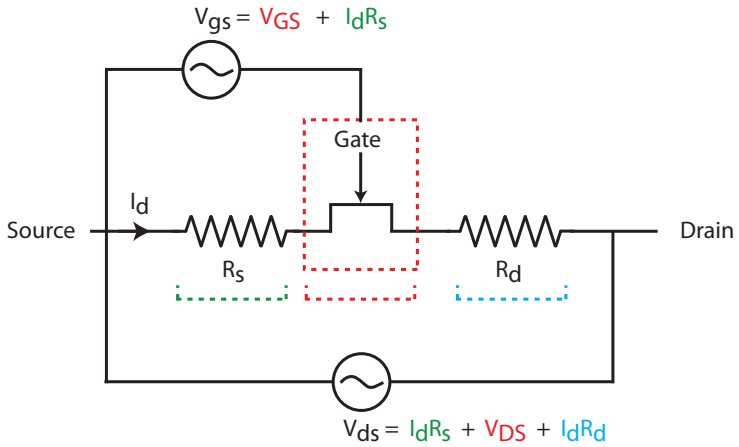


Figure 3.18: HEMT in common-source configuration with parasitic source and drain resistances.

Referring to Figure 3.18, we define the drain-source voltage dropped across the intrinsic region of the device as V_{DS} and the gate-source voltage similarly as V_{GS} (subscript capitals), with the current I_d flowing through the drain-source resistive divider of the device. Hence, defining intrinsic transconductance as $g_{m0} = \frac{\delta I_{ds}}{\delta V_{GS}}$ and expressing the applied gate-source voltage in terms of the voltage drops, the extrinsic transconductance can be derived:

$$V_{gs} = V_{GS} + I_d R_s \tag{3.51}$$

$$g_m = \frac{\delta I_{ds}}{\delta(V_{GS} + I_d R_s)} \tag{3.52}$$

$$= \frac{\delta I_{ds}}{\delta V_{GS} + \delta I_d R_s} \tag{3.53}$$

$$= \frac{\frac{\delta I_{ds}}{\delta V_{GS}}}{1 + \frac{\delta I_d}{\delta V_{GS}} R_s} \tag{3.54}$$

$$= \frac{g_{m0}}{1 + g_{m0} R_s} \tag{3.55}$$

This is equivalent to the short gate length simplification of Equation 3.40, and hence $g_m = WC_s v_s$.

As a consequence, it is clear that the effect of the source resistance is critical to the drive current modulation efficiency and hence the resultant device performance. Large values of R_s will cause the intrinsic and extrinsic transconductances to differ considerably, whilst for a source resistance of zero, they should be identical. It is noteworthy that the drain resistance plays a negligible role, though it will markedly affect the output conductance.

3.7.2 High frequency performance and figures of merit

As the frequency of operation increases into the microwave and millimetre-wave parts of the spectrum, the wavelength of the signal becomes comparable to the dimensions of the circuit and the device itself, leading to transmission-line characteristics of the signal propagation. At such frequencies, the electromagnetic behaviour becomes increasingly complex and dependent on small changes in impedance, significantly altering the magnitude and phase of the propagating signal.

At high frequencies, the circuit can be modelled as a transmission line by Maxwell's equations, but this requires complete self-consistent electromagnetic analysis, precluding the use of this method in real-time measurement of real devices. Two-port network analysis considers the circuit under test as a "black box", whose input and output voltages and currents can be measured at both ports.

Figure 3.19 shows the port configuration of a single HEMT in common-source configuration. Port 1 refers to the input signal, that applied between gate and source, whilst Port 2 is the output, the drain-source current.

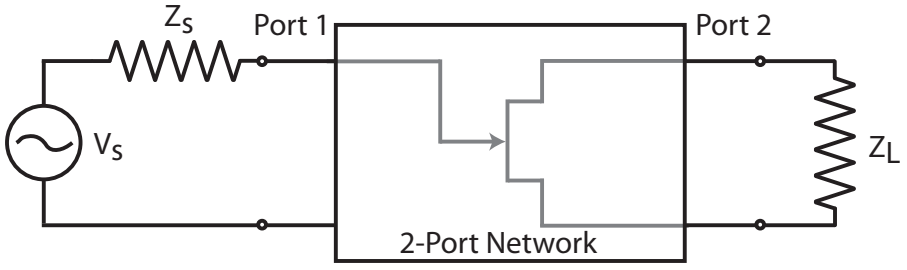


Figure 3.19: Two-port network interpretation of a single HEMT, showing the role of source and load impedances, Z_s and Z_L. The transistor representation will include the extrinsic device.

Considering this configuration, there are several useful figures of merit which can be extracted using two-port analysis. In particular, as with any amplifier system, as frequency increases, gain tends to decrease, defined by the gain-bandwidth product. As the frequency increases, a point will arise where the current gain falls to unity, and where the power gain falls to unity. These are useful figures of merit from the perspective of circuit design, and are referred to as the cutoff frequency and maximum frequency of oscillation, respectively.

Cutoff frequency f_t

At the cutoff frequency, the short-circuit current gain falls to unity, and $i_{in} = i_{out}$. The short-circuit gain is defined as the current gain assuming the output to be shorted, not driving a load. The gain will consequently vary considerably from an optimally matched system.

We consider Figure 3.15 with the output shorted and neglect the small intrinsic resistance, resulting in the circuit of Figure 3.20.

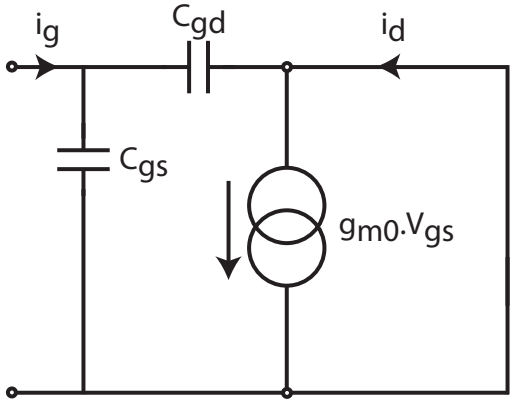


Figure 3.20: Simplified intrinsic HEMT equivalent circuit at short-circuit. The intrinsic resistance, R_i is neglected due to its small relative magnitude.

Considering Figure 3.20, the input and output currents, i_g and i_d, respectively are:

$$i_g = \frac{v_{gs}}{\left(\frac{1}{j\omega(C_{gs}+C_{gd})}\right)} = j\omega(C_{gs}+C_{gd})v_{gs} \tag{3.56}$$

$$i_d = g_{m0}v_{gs} \tag{3.57}$$

The short-circuit current gain is:

$$A_i = \frac{i_d}{i_g} = \frac{g_{m0} v_{gs}}{j\omega(C_{gs} + C_{gd}) v_{gs}} = \frac{g_{m0}}{j\omega(C_{gs} + C_{gd})} \tag{3.58}$$

$$|A_i| = \frac{g_{m0}}{2\pi f(C_{gs} + C_{gd})} \tag{3.59}$$

Hence, for $|A_i| = 1$, the intrinsic cutoff frequency is [83]:

$$f_t = \frac{g_{m0}}{2\pi(C_{gs} + C_{gd})} \tag{3.60}$$

As a consequence, maximising transconductance and minimising the intrinsic capacitances will yield an increased cutoff frequency.

Extrinsic capacitances will add to the capacitance sum to the detriment of f_t [84], whilst the transconductance will be subject to reduction by the source resistance as discussed in Section 3.7.1. Minimising these parasitics will additionally enhance the cutoff frequency. The extrinsic cutoff frequency was hence more rigorously defined as [85]:

$$f_{t_{ext}} = \frac{g_{m0}}{2\pi(C_{gs} + C_{gd})\left(1 + \frac{(R_s + R_d)}{R_{ds}}\right) + C_{gd} g_{m0}(R_s + R_d)} \tag{3.61}$$

To gain a better understanding of the role of the channel electron dynamics in defining f_t, we can also substitute Equation 3.44 into 3.60, such that [65]:

$$f_t = \frac{q v_{sat} \frac{q\delta n_s}{\delta V_{gs}}}{2\pi L_g C_{tot}} = \frac{v_{sat}}{2\pi L_g}\eta \tag{3.62}$$

Hence, the efficiency of the channel electron modulation, coupled to increasing of electron velocity and decreasing gate length, are critical to maximising cutoff frequency. Although this, too, is a simplistic interpretation given non-linear electron transport effects inherent to very short channels, as will be discussed in detail in Section 3.8.2, the general need to increase electron velocity remains true. Coupled to a reduction in gate length, this simplistically implies the relevance of electron transport time along the channel. To this

end, the f_t is frequently expressed by delay-time analysis as [86]:

$$f_t = \frac{1}{2\pi\tau_{eff}} \tag{3.63}$$

Where $\tau_{\text{eff}} = \tau_{\text{transit}} + \tau_{\text{cc}} + \tau_{\text{p}}$. The delays are channel transit time, channel charging time and parasitic charging times respectively [87]. Minimising these times implies an increase in cutoff frequency. Comparing Equations 3.63 and 3.62, it is interesting to note the role of the modulation efficiency in defining the relative lengths of these times.

Maximum frequency of oscillation f_{max}

The maximum frequency of oscillation is related to the power gain of the device while driving a matched load.

The equivalent circuit of Figure 3.15 is modified to model the gate-channel capacitance as a single capacitor for simplicity, with C_{gd} effectively open circuit. The gate resistance is also included from the extrinsic circuit due to its key effect on the input signal.

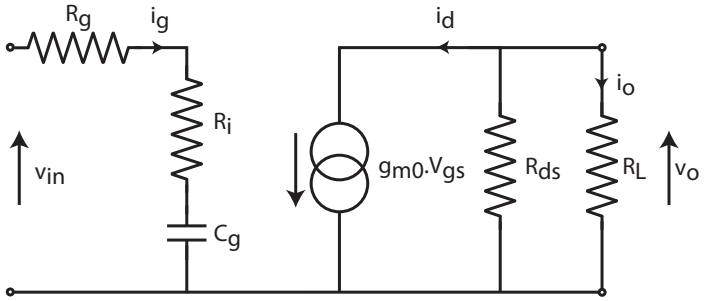


Figure 3.21: HEMT equivalent circuit for driving a matched resistive load. The gate resistance and output resistance are included due to their relevance.

The voltage gain of the circuit is [88]:

$$|A_v| = \frac{v_o}{v_{in}} = \frac{g_{m0}R_o}{\sqrt{1+\omega^2 C_g^2 (R_g + R_i)^2}} \tag{3.64}$$

Where R_o is the output resistance of the equivalent circuit, the parallel combination of R_{ds} and R_L.

Since $\omega^2 C_g^2 (R_g + R_i)^2 >> 1$, and substituting Equation 3.60,

$$|A_v| = \frac{g_{m0} R_o}{2\pi f C_g (R_g + R_i)} = \frac{f_t R_o}{f(R_g + R_i)} \tag{3.65}$$

The maximum frequency of oscillation is defined at the matched load case, when the load resistance equals the HEMT's output resistance, $R_L = R_{ds}$. Hence, $R_o = \frac{R_{ds}}{2}$ and the load receives half the generated current, $i_o = \frac{i_d}{2}$.

Therefore,

$$|A_v| = \frac{f_t R_{ds}}{2f(R_g + R_i)} \tag{3.66}$$

Referring to Figure 3.21, and substituting the definition for f_t in Equation 3.60, current gain is defined as:

$$|A_i| = \frac{i_o}{i_g} = \frac{g_{m0}}{4\pi f C_g} = \frac{f_t}{2f} \tag{3.67}$$

Hence,

$$G_p = |A_i|\,|A_v| \tag{3.68}$$

$$= \left(\frac{f_t}{f}\right)^2 \frac{R_{ds}}{4(R_g + R_i)} \tag{3.69}$$

Setting $G_p = 1$ and rearranging,

$$f_{max} = f_t \sqrt{\frac{R_{ds}}{4(R_g + R_i)}} = \frac{f_t}{2} \sqrt{\frac{R_{ds}}{(R_g + R_i)}} \tag{3.70}$$

The maximum frequency of oscillation is hence maximised by increasing the cutoff frequency as previously described, increasing output resistance and minimising intrinsic and gate resistances.

f_{max} was more completely derived in [81] and [89] as:

$$f_{max} = \frac{f_t}{\sqrt{4g_{ds}\left(R_i + R_s + R_g\right) + \frac{2C_{gd}}{C_{gs}}\left(\frac{C_{gd}}{C_{gs}} + g_m(R_i + R_s)\right)}} \tag{3.71}$$

Although considerably more complex, and still an approximation, the meaning is clear: the reduction of parasitic resistances is key, whilst the ratio of the intrinsic gate capacitances is of particular importance. It is additionally worth noting that the gate resistance, the only equivalent circuit element defined entirely by device processing, is critical to the determination of f_{max}.

3.8 Scaling the HEMT

As already discussed in some detail, gate length plays a vital role in defining HEMT performance. The discussion, however, has thus far focussed on the simple velocity saturation model, considering only the time spent traversing the gate region, as in Equations 3.47 and 3.62. The reality is somewhat more complex, the transport non-linear, as will be discussed in this section.

In addition, one-dimensional scaling of the gate length alone has been considered in depth so far. The effects of scaling, however, in reality are intrinsically two-dimensional, and will also be discussed.

3.8.1 Gate capacitances and resistance

One obvious effect of the reduction of gate length is the linearly correspondent reduction in gate capacitances. Considering the gate and channel as a parallel-plate capacitor,

$$C_{parallel} = \frac{\epsilon W L}{d} \tag{3.72}$$

Where d defines the gate-channel separation, ϵ is the dielectric constant of the material between the gate and channel (though in reality this will be an InGaAs/InAlAs heterostructure), W is width and L length of the gate stripe.

Consequently, the intrinsic gate capacitance will reduce with the gate length, L. As out-

lined in Section 3.7.2, this results directly in increased cutoff and maximum oscillation frequencies. The gate capacitance, clearly, will be far more complex than this simple scenario, due to the effects of fringing electric fields, dopant planes, non-uniform electron densities in the channel and the presence of free carriers or trapped charges throughout the layer structure. The general trend for intrinsic capacitance reduction, however, remains valid.

Less desirably, gate resistance increases linearly with reducing gate length, as Equation 3.73 shows.

$$R_g = \frac{\rho W}{A} \tag{3.73}$$

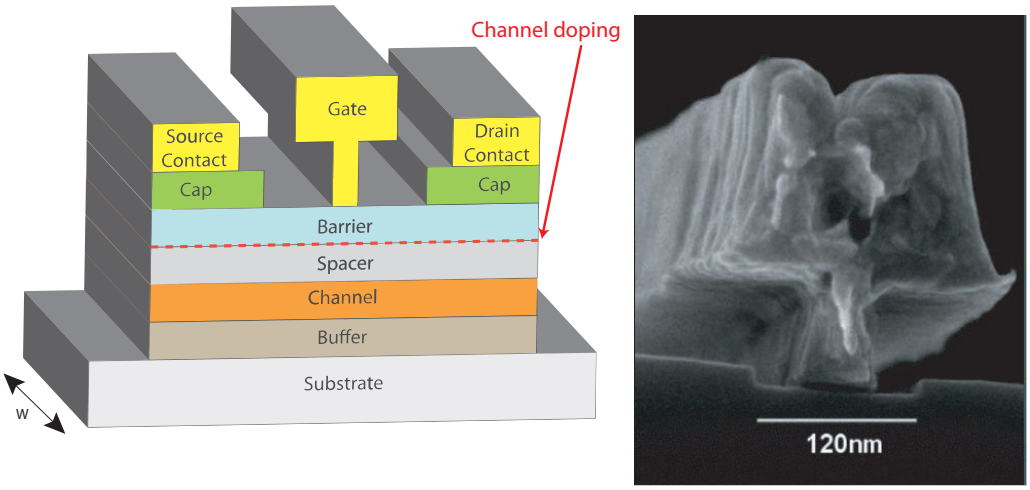
Where ρ is the resistivity of the metal, W is gate width and A is cross-sectional gate area.

As the gate length is reduced, the cross-sectional area reduces linearly. In addition, since producing high aspect ratio metallic structures is problematic from a fabrication technology point of view, described in more depth in the coming chapter, the height of the metallised feature must decrease as the gate length is reduced.

If we therefore assume that the gate foot feature remains cross-sectionally square upon scaling, reducing gate length by half will quadruple its resistance. Considering Equation 3.68, assuming $R_g >> R_i$ and neglecting the transport enhancements due to reduced gate length, the power gain in this case will drop by 75 % and f_{max} will halve (Equation 3.70).

As a consequence, it is crucial to effective device scaling to adopt an advanced gate scaling strategy which does not result in correspondingly increasing gate resistance.

The most common solution is the use of the T-Gate, as shown in Figure 3.22, where the short gate foot is topped by a large gate head, which in reality comprises the bulk of the gate area. As a consequence, it is possible to retain a large cross-sectional area whilst reducing the gate foot length. The gate area can then be controlled by the dimensions of the upper section, such that it dominates the gate resistance. In this way, the electron transport benefits of gate length reduction can be gained without a corresponding sacrifice in gate resistance.



(a) General HEMT schematic showing T-Gate placement. (b) 70 nm T-Gate in gate recess, courtesy of D. Moran.

Figure 3.22: T-Gate layout outline and example SEM micrograph.

3.8.2 Non-equilibrium electron transport

Section 3.6 describes long-channel electron transport, where carrier velocity is governed by low-field mobility and velocity saturation effects. As the gate length is reduced, however, non-linear effects become important due to the short timescales associated with the transit of the intrinsic region.

As previously described in Equation 3.26, the drain current is generally defined as $I_{ds} = Wqnv$, where v is carrier velocity. The magnitude of the drain current is therefore directly dependent on carrier velocity, whilst there is a direct correlation of high frequency performance to electron velocity, as previously discussed.

The term "non-equilibrium" refers to the relative energy imbalance between free electrons and the material in which they are moving. In equilibrium, electrons in a semiconductor have energy equal to that of the surrounding crystal lattice. As electrons acquire more energy as a result of high electric fields, they have greater energy than the thermal energy of the lattice, so are often known as "hot" electrons, as discussed in Section 3.3.2. Electron heating can result in various interesting transport phenomena apart from the low-field, mobility-limited velocity regime.

Intervalley scattering and the relaxation of energy and momentum

As discussed in some depth in Section 3.3.2, III-V semiconductors feature multiple satellite conduction valleys with varying associated effective mass and therefore mobility. Consequently, reduced carrier saturation velocities result in the upper, high energy valleys. Hot electrons can acquire sufficient energy to scatter into the upper valleys from the central Γ valley, resulting in the general velocity-field characteristics of Figure 3.3.

As the electric field is increased, an increasing proportion of carriers transfer to the upper valleys, adhere to their attenuated electron velocities and average electron velocities decrease. Electron velocity therefore intrinsically varies non-linearly with applied electric field.

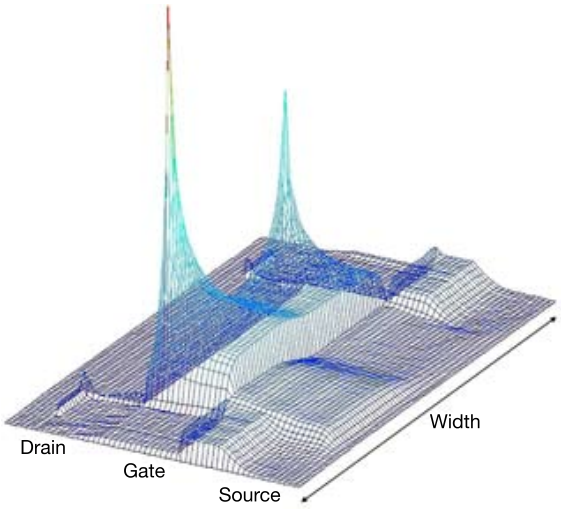
Between scattering events, both momentum and energy are subject to relaxation times as mentioned in Equation 3.7; time is required for each to be lost, governed by the exact scattering conditions of the semiconductor. In general, the two relaxation times are very dissimilar, since the various scattering events can affect energy and momentum differently. Since a high-velocity carrier has a greater probability of scattering [26], momentum relaxation times decrease with increasing field.

Considering firstly the conditions of a single valley, the carrier dynamics become increasingly complicated as these relaxation effects are considered.

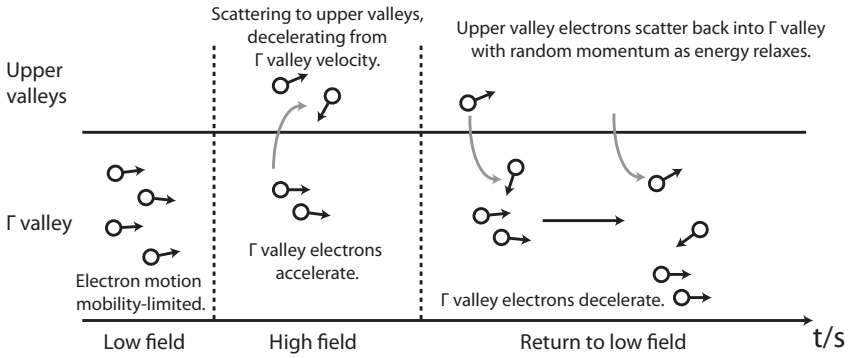
Velocity overshoot

In a HEMT channel, electrons experience a sharply varying electric field as they pass from source to drain as a consequence of the interplay of applied gate and drain bias, with particular peaks in electric field magnitude at the drain end of the gate, since the largest fields result between gate and drain, where there is the largest potential difference as shown in the Monte Carlo simulation of Figure 3.23(a). As a consequence, there is a corresponding variation of relaxation times along the channel. As electrons pass from one field intensity to another, they acquire different saturation velocities.

In the transitional period during which the momentum relaxes, they continue to drift under the influence of the field [31]. As a consequence, an electron moving from a low-field region to a high-field region will decelerate from the low-field saturation velocity, but during the relaxation time will exceed the high-field saturation velocity as a consequence of its excess momentum, as illustrated in Figure 3.24. If the relaxation time is long,



(a) Monte Carlo simulation of field distribution in a 120 nm HEMT. After Kalna, et al. [90].



(b) Velocity overshoot in multivalley semiconductors.

Figure 3.23: Varying electric fields in a HEMT channel and resultant velocity overshoot effects.

therefore, the momentum will take an appreciable time to relax, with a consequently significant distance travelled by the decelerating electron.

Therefore, though the velocity should saturate in high-field regions, over short transitionary distances, this saturation velocity can be exceeded, a phenomenon known as velocity overshoot [26]. Velocity overshoot effects are possible in all semiconductors, but are particularly significant in III-V materials.

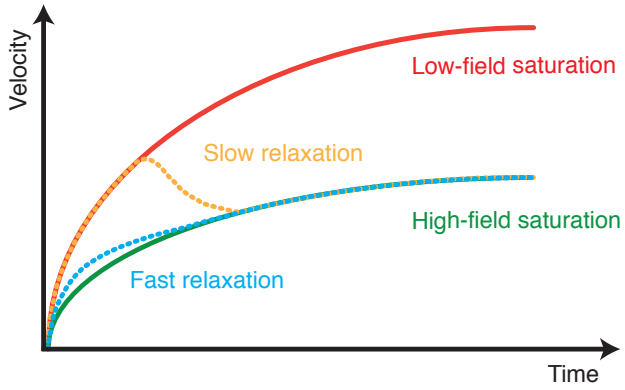


Figure 3.24: Illustration of velocity overshoot with fast and slow relaxation. Adapted from [26]. Long relaxation times equate to low scattering rates.

Multivalley semiconductors such as GaAs or InGaAs have even more complex overshoot characteristics as a result of intervalley transfers as shown in the schematic of Figure 3.23(b).

On reaching a change in electric field intensity, electrons continue to be accelerated by the increased electric field, and can overshoot the high-field saturated velocity. On acquisition of sufficient energy, electrons can be scattered from the Γ to the L and X valleys. Though they can maintain overshoot velocity in the higher valleys, scattering rates in the upper valleys are very high and average velocity decreases more rapidly than in the Γ valley. Moving back into a lower-field region of the channel, electron energy relaxes and these electrons can scatter back to the Γ valley with random momentum; simultaneously, electrons in the Γ valley will also decelerate. The electron velocity will therefore decrease on average, with random contributions from electrons transitioning back into the Γ valley resulting in an average velocity which may dip below its saturation value or even become negative as a result of scattered electrons with negative momentum [26]. Successive valley transfers may also result in greatly increased momentum relaxation times, leading to increased overshoot [91].

As a consequence, over short time periods, hence short distances, velocity overshoot can result in vastly increased electron velocities.

Carriers entering the intrinsic region should do so at a velocity independent of gate length, but dependent on the electric field resulting from voltages applied over a given source-drain separation and recess geometry, and dependent on the material characteristics. Hence, if the effective channel length is reduced on the order of the distances over which velocity saturation occurs, electrons can transit the intrinsic device at the increased overshoot velocities. Increasingly short gate length devices should hence benefit from the effects of velocity overshoot [92, 93].

It is also worth noting that materials with low Γ valley effective mass and large Γ, L and X valley energy separations will incur lower intervalley transfer, with reduced contributions from electrons scattering back into the Γ valley and the highest peak velocities. As a consequence, high-indium InGaAs is favoured as a channel material since the incorporation of indium both increases the valley separations and decreases the effective mass relative to the properties of GaAs.

It is also interesting, as a sidenote, that as the indium content is increased, the probability of impact ionisation increases due to the reduced bandgap, as discussed in Section 3.6. Considering the increased separation between the valence band and the L or X valleys, however, impact ionisation would be reduced for carriers in these high-energy valleys [22]. As a consequence, an interesting situation might involve the velocity overshoot of electrons and their subsequent scattering into the higher valleys whilst retaining overshoot velocity over the intrinsic region. Though this velocity will decrease rapidly as momentum relaxes, carrier energy will remain high for longer and the carriers should remain in the upper valleys. If the gate length was sufficiently short that these carriers did not transit back to the Γ valley and remained at an overshot high velocity, it would be possible to reap the benefits of high velocity carriers whilst minimising impact ionisation-induced phenomena.

Ballistic transport

Velocity overshoot occurs over multiples of the momentum relaxation time, the time between scattering events. The average distance travelled by an electron between such events is known as the mean free path, and is closely related to the mobility as previously noted. The scattering characteristics are summarised in Figure 3.25.

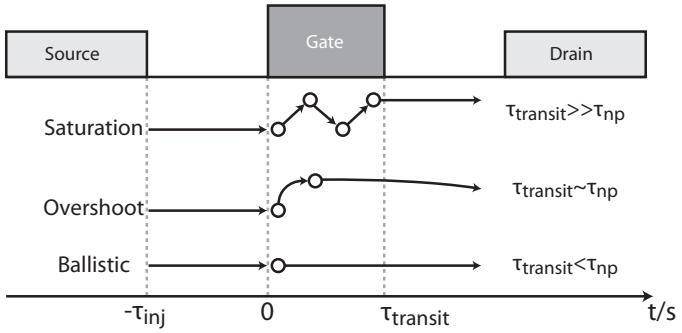


Figure 3.25: Overview of non-equilibrium transport mechanisms, where τ_{transit} is intrinsic channel transport time, τ_{inj} is source injection time and as previously, τ_{np} is momentum relaxation time. Where τ_{transit} is much longer than τ_{np}, velocity will be saturated or mobility-dominated. If the two are comparable, velocity overshoot may be important. If τ_{transit} is less than τ_{np}, transport may be ballistic.

For high-indium InGaAs, 2DEG low-field mobility can be exceedingly high at room temperature, and increases as the temperature decreases. Consequently, the mean free path can be relatively long, from microns at low temperature to tens or hundreds of nanometres at room temperature [94, 95], dependent on the applied electric field.

Over distances shorter than this mean free path, it may be possible for the average electron to avoid scattering and cross the channel unimpeded. As a consequence, mobility becomes meaningless since scattering never occurs, velocity should not saturate and the transport is termed ballistic [94]. As a consequence, a truly ballistic device would feature $\mathrm{I_{ds}} - \mathrm{V_{ds}}$ characteristics in which current did not saturate beyond the linear region. Instead, current would increase with variable conductance as the velocity varied under the applied electric field and electrons would travel with kinematically-defined velocity, v:

$$v = \int \frac{q\xi}{m_e} \delta t \tag{3.74}$$

Where all symbols have their former meanings.

Ballistic electrons would reach an eventual thermal velocity limit imposed by electron injection velocity [96]: the ballistic limit [97]. High indium concentrations act to increase the injection velocity in III-V semiconductors, enhancing the ballistic limit [97, 98]. A ballistic device might be expected to display $\mathrm{I_{ds}} - \mathrm{V_{ds}}$ characteristics similar to those

simulated in Figure 3.26.

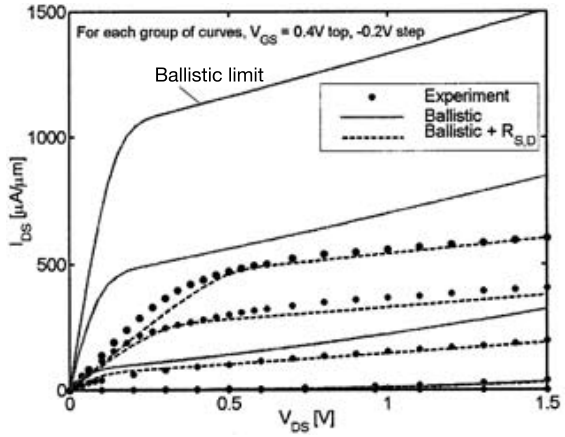


Figure 3.26: Comparison of $I_{ds} - V_{ds}/V_{gs}$ of a fully ballistic transistor from Monte Carlo simulations, compared to a measured 30 nm HEMT. Modified from [99].

As a consequence, the ballistic saturation current would be determined by the ballistic limit itself, $v_{s_{ballistic}}$:

$$I_{s_{ballistic}} = Wqn_s v_{s_{ballistic}} \tag{3.75}$$

At practical operating temperatures, scattering can never be completely ignored as a consequence of the thermal energy of the lattice under high electric fields, which inevitably leads to polar mode phonon scattering [100, 101]. In particular, in multivalley semiconductors, intervalley scattering plays a key role as electric field is increased and carrier energies exceed the valley separation [102]. The terms "near-ballistic" or "quasi-ballistic" are often used to describe this type of electron transport.

It might consequently be expected that the dominant transport mechanism might vary with electric field, with high fields preventing ballistic operation along the device channel. Since mobility is dependent on electric field, the ballisticity of a device channel would vary as the field varies along the channel. As a result, a device which may be ballistic at a medium electric field might be velocity-saturated at higher fields.

Considering an intrinsic HEMT again, as the gate length is reduced, transport should continue to be quasi-ballistic at increasing electric field magnitudes as the gate length becomes similar to the high-field mean free path. As a consequence, reducing the gate

length will increase the probability of quasi-ballistic transport across the complete device, even considering the reduced mobility at high electric field. Device ballisticity is hence limited by the gate length relative to the mean free path associated with peak electric fields in the device [103].

The transfer from diffusive to ballistic transport has been observed in a HEMT 2DEG at low fields [104] whilst other work has verified that the ballistic model is effective for the modelling of an intrinsic HEMT [99]. The parasitic resistances are shown to be the limiting factor in transport, as shown in Figure 3.26, reducing the source electron injection velocity, further emphasising the need for the reduction of parasitic components in addition to the dimensional reduction of the intrinsic device.

An interesting point to note is that whilst a ballistic electron would be immune to collisions, for transport to remain ballistic, the electron must be accelerated to as high a velocity as possible up to the ballistic limit in a time shorter than the momentum relaxation time, during the time τ_{inj} shown in Figure 3.25. The electron acceleration is governed by the application of a high electric field, which consequently reduces the mean free path. As a consequence, for a short-channel device operating ballistically, it is conceivable that electrons may never reach the ballistic limit at all, but remain governed by the electric field, with the injection velocity achieved during τ_{inj} limiting transport.

Less aggressively scaled devices might in fact achieve higher velocities than a ballistic device in this scenario. During overshoot, though electrons are scattered, velocities can greatly exceed the saturation velocity since the electrons have been accelerated by the electric field for significant time.

As a result, it is conceivable that a ballistic scenario might yield lower average velocities than less aggressively scaled overshoot-dominated devices. Average channel transit times may, however, remain shorter as a consequence of the short channel, resulting in improved high-frequency performance despite a reduction in drive current resulting from reduced electron velocity. The ideal target scenario therefore involves retaining fully ballistic transport with the thermal limitations of source injection minimised to maximise the ballistic limit, whilst simultaneously ensuring that the carriers achieve maximum acceleration over a sufficient time.

In practice, real devices are likely to exhibit elements of ballistic and overshoot-dominated transport dependent on the operating conditions, materials and scaling of the complete device.

3.8.3 Effective channel length

Whilst the intrinsic channel region is dominant in defining device performance, it is crucial to note that the actual channel length is in practice somewhat different from the defined gate length. The real channel length will be defined by the region under the influence of the electric fields induced by the gate.

Since the gate and channel approximate a parallel-plate capacitor, fringing electric fields will extend beyond the physical gate length. The result is the extension of the region of the channel under the influence of the gate: the effective gate length of the device [105]. This is often known as the depletion length since it implies the region of the channel that can be pinched off and depleted of carriers as described in Section 3.6.2. This extension effect occurs for all gate lengths, but since the extension can be tens of nanometres, dependent on geometry, it becomes increasingly apparent in short gate length devices where the extension may form an appreciable percentage of the effective gate length. Additionally, the extent to which fringing fields define the depletion length is affected by the aspect ratio of the channel. If the gate-channel separation is large, the fringing fields will extend further, increasing effective gate length. The modulation efficiency will also drop as a consequence of the increased separation.

In the extreme case of extension, the depletion region becomes semi-circular and entirely dominated by the fringing field elements. As a consequence, regardless of the physical gate length defined lithographically, the effective gate length will always remain longer by a finite extension. The consequence of this, considering previous derivations, will be to suppress drive current, transconductance and high frequency performance over that expected for a lithographic gate length.

The physical situation is exacerbated when surface states are considered. In Section 3.5.1, the role of surface states in defining the pinned Fermi level in unpassivated semiconductors was discussed. The effect of surface states, however, is more complex than defining the surface potential, as suggested by their probable role in defining the kink effect, discussed in Section 3.6.2 and generally in Section 3.5. Surface states represent traps and surface charges which can significantly affect electrostatics and the ability to control conduction in the vicinity of the pinned surface. Additionally, these states may have varying corresponding activation energies and response times, with the consequence that their behaviour may vary with bias conditions and operating frequency of a device incorporating unpassivated surfaces.

A normal HEMT layout features a gate formed in a recessed region where the highly-doped cap layer has been etched away. As a consequence, the device barrier layer is exposed on both sides of the gate and surface states are present adjacent to the gate on both sides. These surface states lead to trapped charge in the regions adjacent to the gate, which can be modulated by the gate's fringing fields and causes their extension. The variable charge present at the surface then influences the channel electron density, increasing the effective gate length [106].

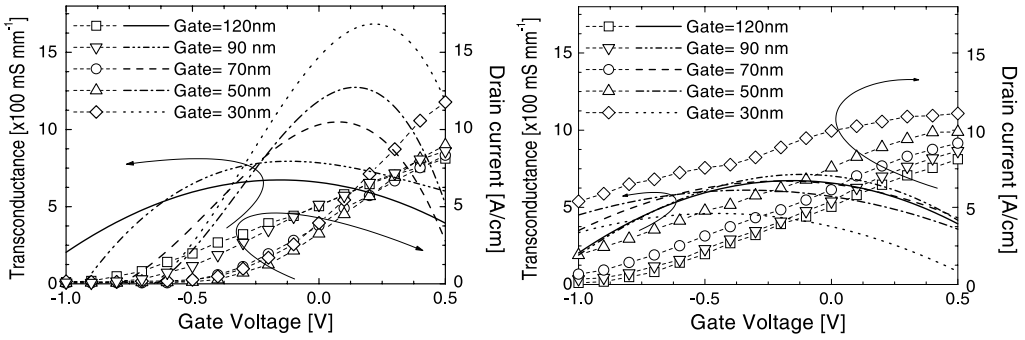
In recent analysis of potential short gate length HEMT performance, Akis, Ferry, et al. [107, 108] defined effective gate length not as the depletion length itself, but the length over which electron velocity is affected by the gate. By Monte Carlo simulation of scaled devices of various gate lengths, the effective gate length was extrapolated for a device of zero gate length, and found to be around 15-20 nm. This work therefore suggests diminishing returns as the device is scaled nanometrically, since the minimum achievable gate length may become much longer than the physical gate length. The role of source-drain scaling into this regime is unclear.

It is also interesting to note that the effective gate length will be determined by the two-dimensional scaling of the device and not entirely by the gate length itself, since the regions under the influence of the fringing fields may be affected by the channel depth [108]. Vertical scaling of device architecture is the focus of the following section.

3.8.4 Vertical scaling

It is critical to note that the HEMT must be scaled vertically as the gate is scaled laterally, since the aspect ratio of the gate length to gate-channel separation is of importance in maintaining the channel electrostatics [103]. A given gate-channel separation may therefore be known to be optimal for a device of a given gate length, yielding both favourable electron dynamics in the channel and the capacity to control the electron population. If the barrier layer is too thin, the pinned surface may have an overly significant impact on the channel population, whilst a thick barrier reduces the effect of the Schottky barrier height variation by the application of a gate voltage. If the gate length is reduced without corresponding epitaxial scaling, the effective electric field induced by the gate voltage acts on a reduced volume of the channel. As a consequence, the channel population is modulated less efficiently and transconductance decreases, to the detriment of potential high-frequency performance.

Proportional vertical scaling of the device architecture to include the thickness of the barrier, spacer and channel is therefore important to maintain adequate channel population control and compensate the effects of the reduced gate length. These effects can be seen in the Monte Carlo simulations of [103], reproduced in Figure 3.27. As is clear, in the fully scaled case, the gate has full control over the drain current for all gate lengths over the bias range, whilst transconductance increases with decreasing gate length. In the laterally-scaled case, transconductance drops off with decreasing gate length and drain current is not easily modulated over the previously well-behaved bias range.



(a) Fully scaled

(b) Laterally scaled.

Figure 3.27: Monte Carlo simulation of effect of intrinsic device scaling on drain current and transconductance. Devices are scaled from 120 nm device dimensions, after [103].

This is a similar situation to the scaling laws that apply for MOSFET gate oxide, where scaling laws act to maintain electrostatic integrity and device current-voltage characteristics between technology generations. Failing to properly scale the epitaxy underlying the gate results in a shift in threshold voltage as well as modulation inefficiency, related to the influence of increased fringing fields from the gate to the channel [109]. As a consequence, the gate-channel separation is generally kept several times smaller than the gate length to maintain electrostatic integrity.

It must be noted that the scaling of the channel thickness is complex in HEMTs, though it is similar to the situation in thin-body MOSFETs. As the channel is thinned, the probability of interface scattering increases as a consequence of the roughness at the boundaries of the channel to the spacer or buffer. By thinning the channel, an electron population of a given density confined within it is therefore more likely to occupy the regions of the channel where the increased scattering has significance. As a consequence, thinning the channel reduces mobility in general. It must furthermore be noted that if

channel transport is highly non-equilibrium in nature, reduced mobility may be irrelevant. The route to an optimally-engineered device architecture for a given gate length is hence increasingly unlikely to simply be a case of maintaining aspect ratio; rather it must consider the complete characteristics of electron transport and population modulation in the device.

It is further interesting that during the scaling process, electric fields under the gate are likely to vary even if scaled according to a "constant electric field" strategy, as a result of the increasingly mesoscopic scale of the intrinsic device. As previously discussed, variations in electric field dominate the transport phenomena, determining overshoot and ballisticity. As a consequence, the complete scaling process will have an effect on channel transport.

3.8.5 Limits to scaling

As the transistor is vertically scaled, the eventual limit to scaling becomes tunnelling through the barrier, which affects both gate leakage and the ability to correctly pinch off the channel. Direct field emission through the barrier is expected to be problematic as the gate-channel distance is scaled to dimensions as small as a few nanometres [108].

Similarly, tunnelling may prove to limit the lateral scaling of FETs as well, as the depletion region may become sufficiently small as to be directly tunnelled by channel electrons by thermionic processes under large electric fields [110, 111]. This is expected to begin to be problematic at gate lengths shorter than 10 nm, though this must be interplayed with the effective gate length argument of Ferry, et al. cited previously [107].

At the ultimate scaling limit, traditional FET geometries will cease to be sufficient as a result of the loss of electrostatic control over the channel [110]. It is therefore expected that fin-gate or double-gate geometries [97, 112, 113] may take precedence in the medium term, whilst an ultimate solution may be a wrap-gate nanowire geometry [114–116]. To date, little work [117] has been done on the high-frequency potential of nanowire structures.

3.9 Summary

This chapter has explored the semiconductor physics behind the operation of the High Electron Mobility Transistor and the device characteristics that result. The advantages

and problems associated with HEMT technology have been detailed, and the requirements for the epitaxial and contact structures explored.

The transport mechanisms of electrons in short channel transistors have been outlined in particular, outlining the benefits of short channel devices and the advantages of increasing the indium composition of the channel. The high-frequency performance of the device has also been investigated and the small-signal RF equivalent circuit examined, as well as the extraction of the key R.F. and d.c. figures of merit. The deteriorative effect of parasitic resistances and capacitances on device performance has in particular been highlighted. The chapter has concluded with a focus on the necessity for and problems with two-dimensional device scaling and the limiting factors therein.

4. Fabrication techniques

4.1 Introduction

The global semiconductor industry has, over the years, driven the growth of myriad essential technologies for the high-yield realisation of circuits which, increasingly, can feature billions of individually complex devices, particularly for digital logic on silicon. As the feature sizes decrease, the control and precision of these processes becomes critical.

As a consequence, developing processes for the realisation of nanometric devices requires in-depth understanding of the techniques used for fabrication. This chapter aims to outline each of the processes used in this work: epitaxial material growth, lithography, deposition processes for metals and dielectrics and etch techniques, whilst ion implantation will be dealt with in Chapter 9. Traditional HEMT process flows are then outlined to illustrate the use of these processes in device fabrication.

4.2 Epitaxial material growth

As discussed in Section 3.4 of the previous chapter, atomically precise control of the device layers is required for successful device operation, particularly in the formation of abrupt heterojunctions or introduction of delta-doping planes. It was the advent of these processes, which arose in the latter three decades of the 20^{th} century, that led to the invention of the HEMT.

The starting substrate wafers are generally produced by liquid-encapsulated Czochralski (LEC) growth methods, where the crystal is drawn from a melt surrounded by a layer of liquid oxide, a method that yields single-crystal growth. The ingots are then sawed and polished into wafers, ready for epitaxial growth. The purity and perfection of the substrate is important, since defects generally propagate during growth [118].

Epitaxy involves the transport of atoms from high purity sources to the substrate wafer, where, under the correct temperature and pressure conditions, layered crystalline compounds are grown in the proportions of the source elements. The two major contemporary methods of epitaxy are metal-organic chemical vapour deposition (MOCVD) and molecular beam epitaxy (MBE), though other methods exist. MOCVD makes use of multiple gaseous sources and organic molecules to transport the atoms to the wafer surface, where the gases react to form compounds with corresponding by-products. It is comparatively low-cost and is frequently used commercially.

MBE provides high levels of interface precision and control over the growth conditions. As a consequence, it is also now widespread and is the technique used for wafer growth in this work.

4.2.1 Molecular beam epitaxy

An MBE reactor consists of multiple effusion cells, arranged around a rotating stage, on which the substrate wafer is mounted, as shown in Figure 4.1.

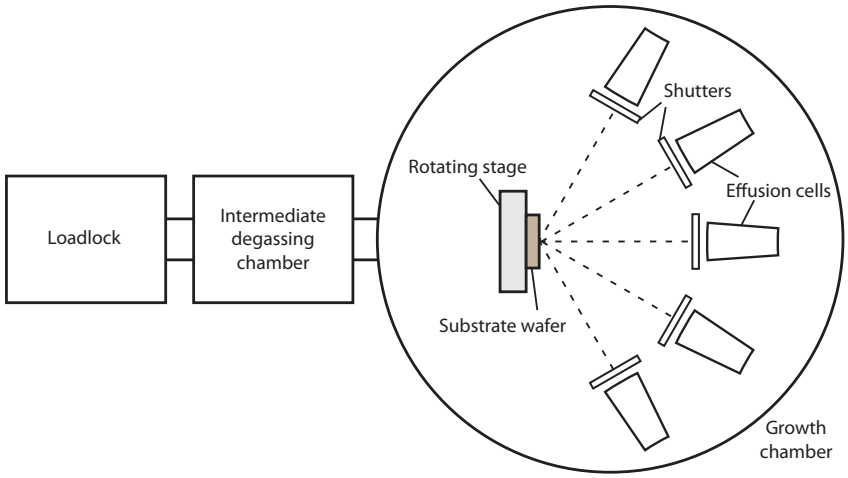


Figure 4.1: Schematic of an MBE reactor.

The complete system is operated under ultra-high vacuum (UHV), typically on the order of 10^{-8} mTorr, which minimises the incorporation of undesirable impurities into the epitaxial layers. An intermediate chamber may be used while loading samples to allow initial degassing and decontamination. The high-purity solid elemental sources are placed in thermally-isolated effusion cells, where they are heated to their sublimation temperatures

of 400-700 °C. The substrate wafer, too, is heated to several hundred degrees Celsius, depending on the growth conditions. Under high vacuum, the vaporised elements have an energetically favourable diffusion route to the heated substrate, where they can be incorporated into a surface film with a given binding rate. Each of the cells has a blanking shutter, allowing the individual elemental flux to be controlled. The relative fluxes of each element are then controlled using the effusion temperatures. This approach allows a very wide range of layer compositions to be grown with atomic accuracy [119] and it is possible to grow highly-uniform single-crystal thin films across a complete wafer.

Issues with MBE include its complexity, cost and throughput, mainly due to its UHV requirement. The system is limited to a relatively small number of wafers per growth session and is a relatively slow technique, requiring at least several minutes per session, often with intrinsic delays to allow ramping and cooling of effusion cells dependent on the required structures.

MBE reactors generally use a reflection high-energy electron diffraction (RHEED) system to monitor the quality of the grown crystal and provide iterative feedback for a following session. This technique operates by the interaction of a beam of electrons with the MBE-grown surface, in order to determine the crystalline nature of the material. An electron beam collides with the surface at a shallow glancing angle, whilst the semiconductor lattice forms a diffraction grating. The diffraction pattern produced allows a determination of the crystalline quality, be it single-crystal, polycrystalline or amorphous, and the presence of impurities. A perfect crystal forms a diffraction pattern consisting entirely of lines, separated by a distance proportional to atomic spacing. Rough surfaces have additional interference effects, whilst amorphous material results in ring-shaped diffraction patterns. Modern MBE systems increasingly employ *in situ* RHEED to monitor growth conditions, since the diffraction pattern intensity varies during monolayer formation as roughness increases, then drops on completion of a monolayer. The diffraction grating should therefore consist entirely of lines on completion of a layer, with additional patterns varying in intensity during layer growth [120].

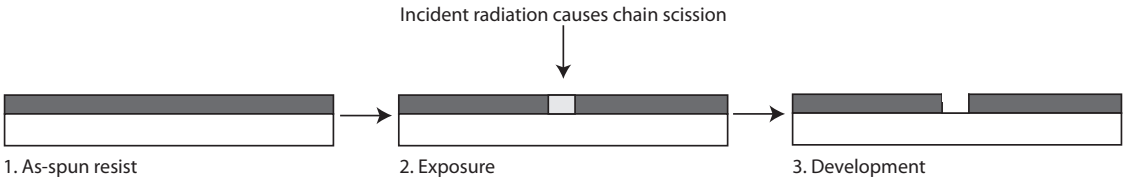
The MBE reactor used in this work is a Varian Gen III dual-chamber machine, which also has an interlocking oxide growth chamber. The main chamber features gallium, arsenic, aluminium and indium, with silicon and beryllium as dopant elements, whilst the oxide growth chamber has cells containing polycrystalline Ga_2O_3 and In_2O_3, elemental gadolinium and oxygen and nitrogen. As a result, a wide variety of III-V layers can be grown along with useful oxides for III-V MOS applications.

4.3 Lithography

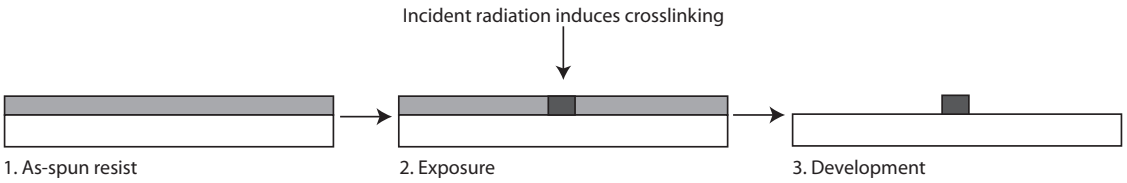
Lithography is the process of patterning a surface using a masking layer, and is the cornerstone of modern semiconductor fabrication.

The term "lithography" encompasses many techniques, including "soft lithography" such as stamping or printing using soft pattern transfer stamps [121] or imprint methods [122]. In particular, the nanoimprint technique has received much recent attention [123, 124].

Most modern methods of lithography, however, employ a radiation-sensitive film, usually a polymer, referred to as resist, which is uniformly applied to the substrate by spin coating or spraying. On radiation exposure, the chemical structure of the resist changes, becoming either more or less soluble in a given developer solvent.



(a) Positive resist.



(b) Negative resist.

Figure 4.2: Positive and negative resist development.

In so-called positive resists, irradiation causes depolymerisation by chain scission [125]. Scission event probability is proportional to irradiation time, so longer exposure times break down the polymer film further. During the development process, an appropriate developer can then dissolve the depolymerised regions, resulting in their removal given a sufficient dissolution time. In positive resists, therefore, exposed regions become unmasked.

In negative resists, the unexposed resist is soluble in a given developer. Exposure results in radical cross-linking of the polymer chains, increasing their density and average chain length. During development, these exposed regions are relatively insoluble and unaffected

by the dissolution process, whilst the unexposed regions are removed. Negative resists, therefore, mask the substrate only in exposed regions. This can be extremely useful for situations where large areas of material need to be removed since the exposure area is reduced.

Figure 4.2 outlines the lithography process in cross-section for positive and negative resists.

The achievable resolution of a lithographic system is therefore defined by several factors: the source and type of incident radiation and the chemistry of the resist exposure process. The radiation used must hence match the resist film sensitivity, and defines both and the equipment and techniques used during exposure.

4.3.1 Optical lithography

The two main demands on lithography in CMOS fabrication are for high throughput to provide low unit processing costs, married to ever-decreasing minimum feature sizes. As a result, optical lithography dominates most pattern definition in the semiconductor industry, thanks to its low running costs and relative simplicity coupled to extremely high throughput. Capital costs are usually astronomically high to achieve these goals with high yield. Optical lithography relies on the incidence of UV light onto the resist-coated substrate through a hard mask, generally defined by electron beam lithography, which, as a core technology used in this work, will be described in detail in Section 4.3.2.

Basic photolithography uses a mask held in contact with the sample to be processed using a vacuum, with UV light focussed through a lens to ensure coherence and uniformity of exposure. This system, known as contact photolithography, ensures optimal pattern transfer from the mask, but as a result of the physical contact, carries a great risk of mask damage and sample contamination. As a result, the mask is generally elevated above the substrate, which, though it removes any potential for damage or contamination, degrades the achievable resolution as a result of diffraction around mask features. The minimum achievable resolution, l, is defined by the product of the wavelength of the light, λ, and mask separation, s:

$$l = \sqrt{\lambda s} \tag{4.1}$$

As a consequence of the resist thickness, combined with a wavelength of 193nm (com-

monly produced by an ArF excimer laser and known as deep ultra-violet (DUV)), contact lithography could theoretically achieve a resolution of around 300 nm from 500 nm-thick resist. In practice, however, most laboratory contact lithography systems have a minimum feature size of around 1 μm as a result of increased mask offsets and their use of cheaper 365 nm light sources.

Industrially, the feature sizes have been substantially improved over this, as a consequence of the use of a reduction lens between the mask and sample, known as optical projection lithography, now common in modern optical steppers [126].

The resolution for a projection system is then given by [127]:

$$l = \frac{\lambda k}{NA} \tag{4.2}$$

Where k is a resist-related ideality factor, and NA is the lens numerical aperture, which specifies the refractive properties including aberrations. NA is a number between zero and one and is in effect a measure of the angular extent of the lens. The use of modern high-NA lenses and low-k resist systems has resulted in optical stepper systems which are capable of resolutions as small as 30 nm [128]. Using multiple exposures [127] and phase-shifting techniques [129], as well as lateral resist etchback [130] or deposition-driven processes [131, 132], the silicon industry has continued to use optical lithography as feature sizes have shrunk to well below the apparent physical limitations of the light source used; a decision driven by both cost and throughput. Future lithographic efforts in the silicon industry look set to continue this route, the current best options for extending optical lithography being the use of short-wavelength extreme UV (EUV) sources [133–135] and immersion techniques [136, 137], where the air gap is replaced by a liquid of high refractive index, extending the resolution by the same factor. Though these methods promise increased resolution, the challenges for both are immense. In particular, since air significantly absorbs light at shorter wavelengths, EUV use requires vacuum conditions, and is hence extremely expensive. The combination of these technologies is, however, expected to meet the immediate requirements of the 32 nm ITRS logic generation and below [138, 139].

Optical lithography, however, in addition to its wavelength-limited resolution, has a number of other drawbacks. A modern optical stepper system is immensely expensive, partly due to the virtual perfection of the required optics, and also as a result of their

shallow depth of field, which decreases with increasing NA [126]. As a consequence, mechanical and thermal stability and guaranteed process repeatability are of paramount importance. As a result of these factors in addition to associated new materials and techniques, a new ITRS process iteration requires fabrication facilities whose capital costs are in the region of billions of dollars; exclusive to a handful of companies globally.

4.3.2 Electron beam lithography

Electron beam lithography, often referred to as "e-beam" or EBL, uses a focussed beam of electrons to write patterns directly onto the resist, with no need for a mask as in optical lithography. Instead, the patterns to be written are generated in software and transferred directly to a computer-controlled exposure system which controls the highly-confined electron beam to produce the pattern.

The primary advantage of using electron exposure is the removal of the resolution constraints inherent to the diffraction of light. Diffraction in particle beams is a consequence of their wave nature, limited by the de Broglie wavelength of an electron beam, which at 100 kV accelerating voltage is around 4 pm: smaller than the atomic spacing of any substrate material [140, 141]. As a consequence, electron beam lithography systems are not generally limited by diffraction; instead the minimum achievable pattern dimensions are driven by the spot size achievable by the electron optics and the interactions of incident electrons with the resist and substrate during the patterning process.

Column layout

Figure 4.3 shows the layout of a typical electron beam lithography system with a thermal field emitter (TFE). Electrons are emitted using the TFE process outlined in Section 3.5.2, using a zirconium oxide-coated tungsten cathode, which is heated and biased to emit electrons omnidirectionally. The column high-voltage (HT) source can usually bias the cathode up to accelerating voltages of several hundred kV. Higher accelerating voltages provide better emission collimation, giving smaller spots at the expense of smaller current densities, hence slower write times.

The suppressor electrode surrounds the emitter such that electrons are only emitted from the cathode tip. A second electrode, the extractor, is used to create a large electric field between it and the cathode, allowing thermionic field emission from the source and electron acceleration towards the extractor. An additional focussing electrode further

confines the electron beam before it reaches the anode.

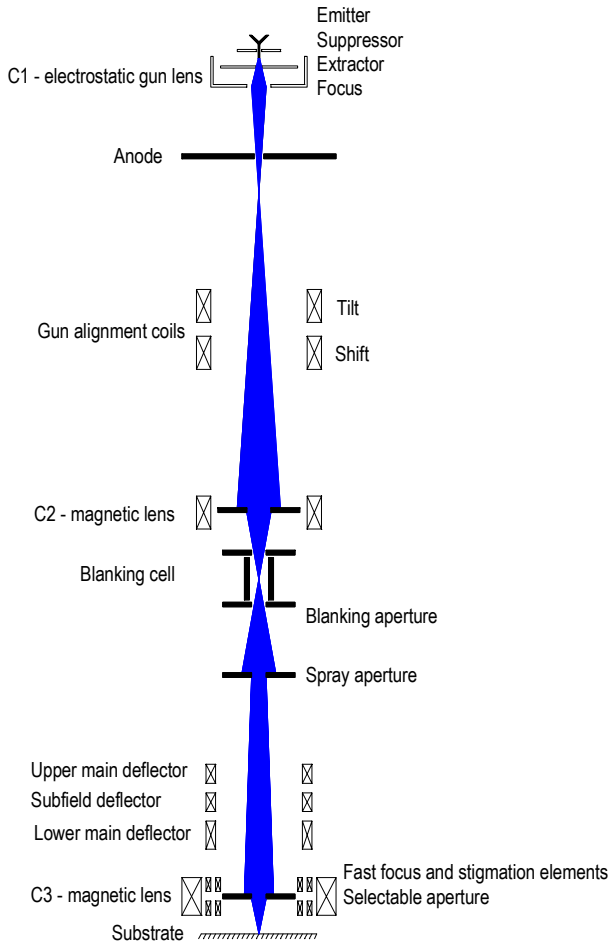


Figure 4.3: Schematic of an EBL system thermal field electron emission column. After Vistec Semiconductor Systems, Vectorbeam operator manual [142].

The electron beam is focussed using a series of magnetic and electrostatic lenses. The gun alignment coils allow alignment of the electron beam to the central two-dimensional axis allowing optimal spot formation and focussing by the lenses. Various apertures are also located at points down the column, which act to constrain the beam divergence to produce a given spot diameter, with a corresponding trade-off in current density.

The lenses themselves comprise the electrostatic lens at the source, marked C1 in Figure 4.3, whilst further magnetic lenses, C2, allow zooming of the spot diameter without changing the beam current and focus. The final magnetic lens, C3, acts to define final

beam focus for a given working distance, and hence the quality of the final lens strongly influences the spot geometry and focus.

The spot diameter on the substrate, d, is determined by [140] :

$$d = kC_s^{\frac{1}{4}}\lambda^{\frac{3}{4}} \tag{4.3}$$

Where C_s is the spherical aberration of the final lens, λ is the electron wavelength and k is an ideality constant that defines the coherence of the beam current to its diameter, effectively determining the edge sharpness of the spot. Broers, et al. [140] have calculated the minimum possible ideal spot size to be around 0.37 nm at 100 kV.

The final lens assembly generally also includes the final double-quadropole coils for the correction of beam astigmatism that results from axial misalignments or optical aberrations.

In practice, modern e-beam tools have minimum spot sizes in the region of 1-5 nm, as a consequence of imperfect electron optics.

The beam, crucially, can be deflected using the series of coils above the final lens, whilst a beam blanker allows the substrate to be masked from the beam after a desired period of time. The deflection coils allow the beam to be steered electromagnetically around a fixed area, allowing selective exposure within that region without any movement of the substrate relative to the column.

The beam can only be deflected by a fixed maximum distance before the beam current density and spot shape becomes non-uniform. The column features two beam deflection systems, the sub-field deflection coils which have short actuation times and smaller maximum deflections, and field deflection coils, which, though slower, have a larger range. As a result, a large-area pattern is written by splitting the pattern into many smaller fields. A pattern is first split into trapezoidal sub-shapes known as subfields, with areas within the maximum range of the faster subfield deflectors. A 64×64 array of subfields then comprises a field, whose area is determined by the maximum deflection of the field deflection coils and the pattern generator resolution. The tool used in the department has a maximum field length of around 1.2 mm at a pattern resolution of 1.25 nm, determined by the 20-bit addressing accuracy of the pattern generator, corresponding to a subfield size of around 20 μm. At smaller pattern resolutions, the areas of both field and

subfield are smaller [142].

The undeflected beam is initially placed in the centre of a subfield specified by the pattern generator. The subfield deflection coils are used to control the spot position to uniformly expose a complete subfield. The field deflection systems are used to deflect the beam to the next subfield, addressed by the pattern generator and the subfield deflection coils used to expose that subfield. This process is then repeated until the complete field is written. The use of slower field deflection systems in conjunction with fast subfield deflectors allows the bulk of the beam deflections to be done using the faster systems, with the slower systems required a maximum of only 4096 times per field.

If the patterns to be written are larger than a single field, several fields are stitched together in the same way, but stage movements are required to reposition the beam in the centre of a new field. A precision translational stage must be used, usually driven piezoelectrically and controlled using feedback of the stage position generally determined using laser interferometers. The vertical offset between the lens and sample surface is determined using laser reflectometry. Since the stage movements rely on mechanical precision, errors in its position are generally much greater than those of the beam deflection coils. This is a key EBL performance metric, known as field stitching accuracy.

It is critical to note that EBL systems have very small depth of field, hence correct focus is crucial. Since samples have varying thicknesses, the measurement of the offset to the lens is important for correct focussing, which is then achieved using closed-loop feedback control of the stage offset to the final lens using the reflectometer signal.

A Faraday cup is also positioned on the stage to allow a measurement of beam current density.

Control of the exposure is achieved using beam blanking and relative movements of both the spot and stage; the dose administered is set by the dwell time of the beam on the substrate and its current density. An exposure control computer then controls the beam deflection, blanking, corrections and stage movements to realise a complete pattern previously defined in software.

The beam formed by the EBL tool is either a simple spot of a given size or a predetermined shape realised by apertures. The former is known as a Gaussian beam system, the latter a shaped beam system. Shaped beam systems are frequently used in industry, whilst most research tools have Gaussian distributions. The Gaussian beam profile is

determined by the electron optics, producing a non-ideal Gaussian energy distribution, rather than the idealised delta function of a perfectly collimated beam. This effect, together with the effect of the spot shape, can be compensated in the exposure process, such that the spots overlap to yield uniform exposure. This process is described in the following section, and is outlined in Figure 4.4.

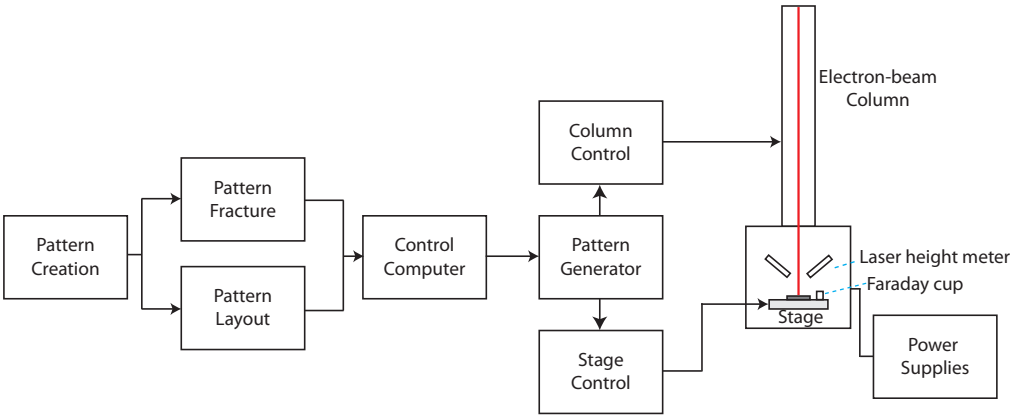


Figure 4.4: General outline of the pattern generation and exposure control of an EBL system.

Pattern definition

Patterns to be written by EBL can be created by standard methods using CAD packages. Since the lithography tool writes large patterns from conjoined small fields, however, the large area structures must be split into shapes realisable by the pattern generator. To achieve this, the fracture process breaks large shapes into trapezoidal subshapes which can be realised within a subfield and potentially into larger fields for stitching. The general process from design to lithography is outlined in Figure 4.4. After the CAD stage, usually resulting in a GDS layout file, commercial fracturing tools such as the CATS package from Synopsys are used to produce the fractured pattern files. The layout of these fractured files can then be specified using a command file which instructs the control computer to expose the fractured patterns at given positions relative to substrate corners at given exposure doses. At Glasgow, these files are created using the Belle software created by Stephen Thoms.

The EBL control computer then reads these command files once the substrates to be written have been loaded by an operator. The pattern files are then read by the pattern generator and exposure times calculated using the dose figures specified in the command

file. The pattern generator then calculates beam positions and dwell times and instructs field and subfield deflections and stage translations accordingly to realise the complete pattern at the intended dose.

Using metallic or topographically etched markers produced in an initial layer of lithography, further layers can be aligned relative to the first to produce multi-level devices. This is generally done by scanning for the edges of a marker of a specified size relative to an easily-found large feature at a known substrate location. The electron beam can also be used for imaging using a backscatter detector, as in an SEM, allowing operators to image the sample for starting lithography operations. Backscattered electron profiles are also used in the marker edge locate routines; hence metals with a large backscatter coefficient contrast to the substrate are required, or a sufficiently sharp etched edge to give backscatter contrast.

All electron beam lithography for this work was carried out using a Vistec VB6-UHR-EWF, which has a maximum field size of 1.2 mm, minimum resolution of 0.5 nm and a minimum spot size of approximately 4 nm.

Scattering and the proximity effect

The most significant limitations to realisable feature size in EBL are due to electron-resist and electron-substrate interactions. As electrons penetrate the resist, they undergo scattering, randomising their velocities. This scattering significantly increases the lateral exposure profile in the resist. Scattering processes can also produce secondary electrons, which then travel at randomised directions with low energies. These low-energy electrons therefore generally have a small range. Figure 4.5 outlines the scattering processes at work in resist exposure.

Scattering that results from interactions with the resist as the electrons penetrate towards the substrate is known as forward scattering. As the incident electrons cause polymer scission, some energy is transferred, scattering the electrons. Higher-energy electrons incur less forward scattering, hence the effect is most significant at low voltages. The effect of forward scattering is to broaden the Gaussian beam energy profile in the resist [140].

Electrons also penetrate the substrate to a depth proportional to the accelerating voltage during exposure and are scattered in random directions. Those scattered with high velocity towards the surface can re-enter the resist, exposing it. Since these electrons

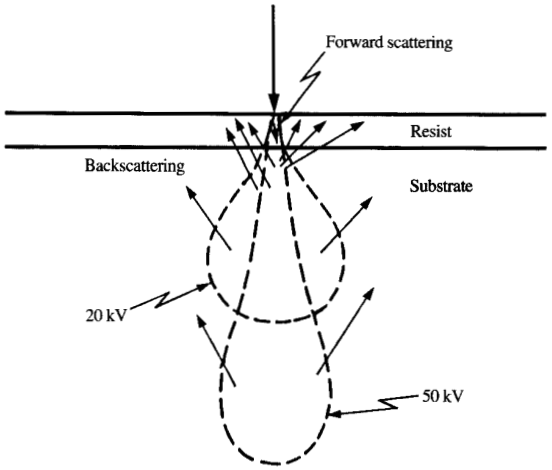
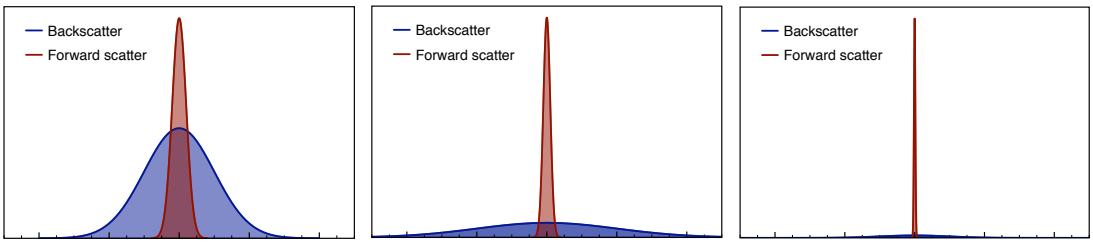


Figure 4.5: Schematic showing interactions of incident electrons with resist and substrate. After [140].

expose the resist from below, the process is known as backscattering. Some electrons are backscattered with sufficient energy to be re-emitted from the surface; indeed this process is frequently used for electron imaging. Different substrates backscatter electrons to a greater or lesser extent. The effect of backscatter is to produce a second Gaussian energy distribution, much broader than the incident beam [140].



(a) Low kV. (b) Mid kV. (c) High kV.

Figure 4.6: Schematic of variation of forward scattering and backscattered exposure contributions with accelerating voltage, where the x-axis depicts lateral energy distribution and the y-axis the relative energy densities. Adapted from [140]. Scales are arbitrary but identical for all figures.

The total exposure profile is therefore a double Gaussian shape, with the backscatter profile superimposed on the forward beam energy. The relative profiles, again, are accelerating voltage-dependent, as shown schematically in Figure 4.6. Higher voltages imply reduced forward scattering, but also a reduced contribution from backscatter, since the

backscattering radius increases for electrons with increased energy. The effect of this is to broaden and therefore smooth the backscatter curve, such that the background exposure from backscattering is broadly uniform relative to the energy of the forward distribution [140, 141].

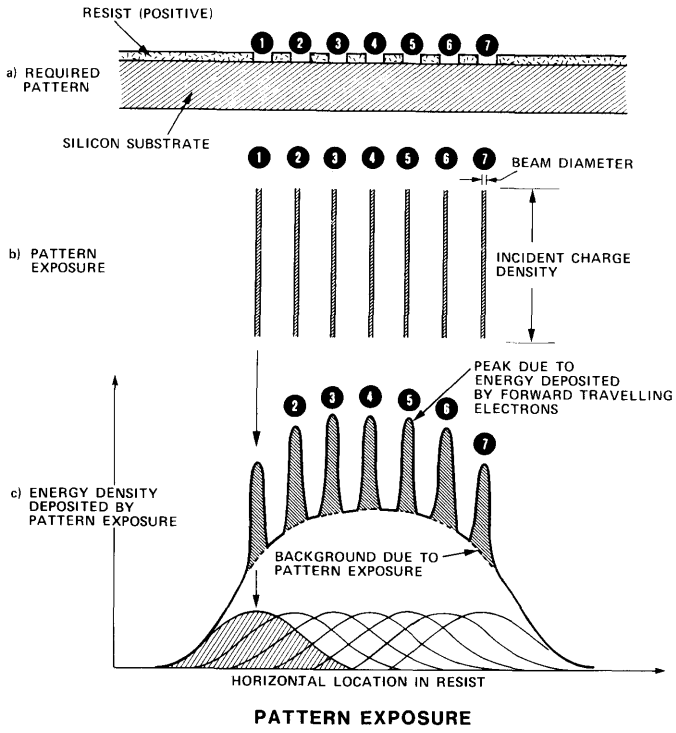


Figure 4.7: Origin of the proximity effect: effects of forward and backscattered electrons on total exposure profile. After [143].

The proximity effect is related to electron backscattering [144]. When writing patterns, as a consequence of backscattered electrons in the resist, densely-patterned regions will receive a higher effective exposure than isolated regions [141]. As a result, when writing almost all multiple-spot features, areas around the edges will receive a relatively lower dose than the centre of the pattern. The effect, shown in Figure 4.7, is due to relative scattering contributions during exposure. As a result, when patterns are placed together, doses required to develop out the corners of the structure result in overexposure of the edges. This is a particular problem when writing two wide features in close proximity; as is the case when defining HEMT ohmic contacts.

The solution is to assign the dose dynamically based on the pattern density or size [145, 146] or to modulate the exposed pattern dimensions accordingly. This can be

done using specialist software at the pattern fracture stage, an example of which is the Proxecco [145] software from Synopsys, which is built into the CATS fracture suite.

Resists and development

The resist used in lithography has a significant impact on the pattern definition process. Different resist chemistries affect developer requirements, etch resistance and exposure characteristics. As a result, some resists have superior resolution, whilst others yield improved etch resistance, solubility or adhesion.

The development time and temperature strongly influence pattern transfer, and have an interaction with the exposure dose. In general, raising the developer temperature increases resist solubility, resulting in faster pattern development. Developing for longer dissolves an increased exposed resist volume. Other process conditions such as agitation during development, can also significantly affect the characteristics [147]. During the development process, the exposed resist is dissolved. There is, however, a time delay associated with the chemical processes of development and development is not immediately complete on introduction into the solution. As a result, developing for too short a time will leave a portion of the exposed resist on the substrate, whilst developing for too long will further erode the resist exposed by backscattering, resulting in enlargement of the pattern. There is therefore an optimal development time for a given exposure dose.

As a further complication, the contrast achieved in the resist between exposed and unexposed areas is dependent on both exposure dose and development time, in addition to the developer and resist system used, as a consequence of the relative dose contributions from forward and backscattered electrons. Different resists and developers yield different contrast characteristics, which may be measured by obtaining the contrast curve for the system in question, which measures residual resist for a given experimental setup. Within a development setup, however, different combinations of exposure doses and development times can affect the resultant contrast.

Resist sensitivity is also relevant, since more sensitive resist requires a lower dose or shorter development times for a given application, reducing processing times and overheads. The sensitivity, as well as etch resistance amongst other variables, is determined by the molecular weight of the resist: effectively a measure of polymer density [148]. Increased molecular weight implies reduced sensitivity as a consequence of the increased required number of scission events for exposure. Molecular weight also affects the ulti-

mate resolution of a resist [149].

Resist film thickness is also a key variable, since thicker films require larger doses and longer development times. Resist films are usually spin-coated, where the spin speed, acceleration and coating time determine the resultant film uniformity and thickness. Resists are usually dissolved in a casting solvent such as o-xylene or di-chlorobenzene, driven off following application by a pre-exposure bake, either in an oven or on a hotplate. It is usual to perform dose testing for a new structure before it can be used, to ensure correct pattern transfer.

The most common positive electron-beam resist is poly-methyl methacrylate (PMMA), which has excellent resolution and is easy to process. PMMA is available in a variety of molecular weights and has a wide processing range [149]. Various co-polymer resists also exist for PMMA where methacrylic acid is added. PMMA-derived resists are generally developed in methyl isobutyl ketone (MIBK) or methyl-ethyl-ketone (MEK) [150], diluted in isopropyl alcohol (IPA).

A second family of unamplified positive chain scission resists is the novolak type [125]. This type combines high-sensitivity with high etch resistance realised by adding novolak polymers. From this basis were developed poly(methyl-chloroacrylate-co-methylstyrene)-based resists, which also produce high etch resistance with reasonable sensitivity. The ZEP series of resists from Nippon are an example of this type. Of particular note is ZEP520A, which retains excellent resolution and contrast with high dry etch resistance [151, 152]. ZEP is usually developed in o-xylene or an alkyl-acetate developer [153, 154].

Most negative resists are based, conversely, on cross-linking of the polymers. An example of an unamplified negative resist is hydrogen silsesquioxane (HSQ), which is a flowable inorganic oxide. HSQ is based on silicon dioxide with available SiH bonds, enabling cross-linking under the provision of energy [125]. HSQ provides excellent resolution and etch resistance to dry etch conditions not designed for silicon dioxide etching [155–157], but has a major problem; it is very difficult to remove, since the oxide formation is permanent. The most common type is the FOX series from Dow Corning.

Chemically-amplified resists

A newer generation of resists is the chemically-amplified type (CAR). CARs operate differently from traditional resists; incident electrons result in the formation of acids in the resist chemistry. These acids then catalyse reactions which occur during development in

the exposed regions [158]. A post-exposure bake is always required before development of a CAR, allowing the acid to diffuse throughout the exposed resist and to provide the required activation energy for reaction. CARs therefore have several additional variables; post-exposure bake time, temperature and delay times between exposure, bake and development. In particular, the sensitivity and contrast are affected by bake and delay times. The environmental variables are therefore critical during processing. CARs have the general advantage of improved sensitivity and, in some cases, contrast.

An example of a positive-tone CAR is the Shipley UV series, which is a copolymer of hydroxy(styrene) and t-butyl acrylate. Originally designed for DUV optical lithography, some are also electron-sensitive, such as UVIII. It has also found use in EBL as an extremely high-sensitivity, etch-resistant resist [159].

One negative-tone CAR is the NEB series of resists from Sumitomo, based on poly(p-hydroxystyrene) [125]. An example is NEB11, well known for its high resolution and extreme resistance to most dry etch chemistries. As with most other negative resists, it is very hard to remove in solvents after development.

A wide variety of resists was used in the course of this project as shown in Table 4.1.

Resist	Type	Application	Relevant section
PMMA	+ve	General EBL	4.6, 7.7
ZEP520A	+ve	High-resolution gate lithography & etching	7.4
LOR	+ve	Undercutting for T-gates	7.2
UVIII	+ve CAR	T-gates	7.2
NEB31	-ve CAR	Etch & implant masking	9.3

Table 4.1: Resists used in the project for various applications.

4.4 Pattern transfer

Any form of pattern transfer is either additive or subtractive. An additive process relies on the deposition of material to the lithographically-exposed areas, whilst a subtractive process removes the current material exposed using etch processes.

All pattern transfer processes, particularly when used on semiconductor substrates, have the capacity to degrade the electronic performance of the underlying material. These processes are often not well-understood, but may be related to an increase in trapping centres with a corresponding decrease in carrier density. For example, etch processes

on GaAs have been shown to degrade the transport properties of the GaAs measured before and after processing [160]. In such cases, damage is often understood as a change in Fermi level pinning at the surface, creation of trapped charges or interface states [161, 162] associated with a "damage layer" of a given thickness [163]. Damage may also be understood as physical changes to the semiconductor material, by either undesirable removal or addition of material. In addition to increasing surface roughness, which may intrinsically be a problem, these changes may alter the surface charge, affecting the electrical properties.

As a consequence, the management of damage to the semiconductor layers is crucial when considering pattern transfer processes.

4.4.1 Additive and subtractive processes

The definition of a metallic contact can be achieved by either additive or subtractive processes. In a subtractive solution, the metal film is deposited onto the sample surface, often by sputtering, then a negative resist is usually used to mask the areas in which the contacts will be formed. An etch process then removes the metal everywhere not covered by the resist. Dry etch processes are commonly used to pattern refractory metals in the silicon industry, though this is less common in III-V.

The additive alternative is lift-off of metal into an exposed region. Using this technique, windows are defined in a double layer of positive resist where the contacts will be formed, using a more sensitive resist at the bottom of the bilayer. If the resist sensitivities, exposures and development processes are correctly chosen, the resultant resist profile should be undercut, with the bottom layer more developed than the upper. As a result, using a directional evaporation process, the deposited metal film will be discontinuous at the edges of the exposed resist areas. In the exposed regions, the metal is deposited on the substrate; in the masked areas, the metal is deposited on the resist. By then soaking the sample in solvent, the resist can be dissolved, removing the metal deposited on top of the unexposed regions. Figure 4.8 shows this process. In general, a lift-off process requires that the bottom layer of resist be as thick as the metal to be deposited.

Both process flows have their advantages. The subtractive approach yields better uniformity and adhesion with correspondingly improved yield, and is generally a "cleaner" approach than lift-off. An additive approach, however, virtually eliminates the possibility of process-induced damage to the underlying semiconductor, since the unexposed bulk

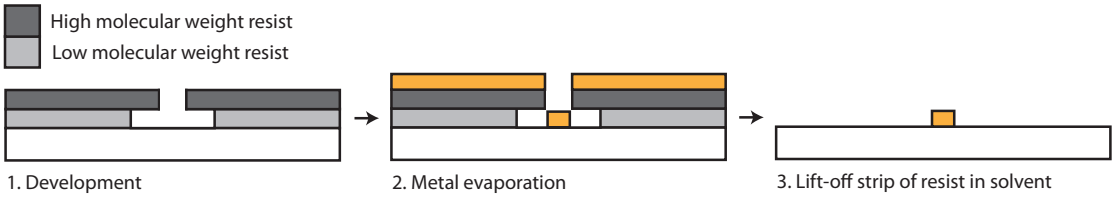


Figure 4.8: Schematic of the lift-off process for evaporated metals.

of the substrate is not exposed to either the metal deposition or etch processes, and is considerably cheaper than the subtractive patterning approach, with its requirement for more complex tools.

Despite the yield advantages of a subtractive approach and its widespread adoption in the silicon industry, most III-V process flows rely on lift-off approaches to metallisation as a result of their ease of damage-free processing and low capital costs. The following sections deal in depth with the physical processes of etching and deposition.

4.5 Material deposition and removal

Selective deposition of additional materials forms the bulk of device processing, using pattern transfer techniques discussed in the previous section. Various techniques exist for metal and dielectric deposition and etching, dependent on the materials and characteristics required. In general, most devices will require metal deposition to form ohmic or Schottky contacts, and most will additionally require some patterning of dielectric films, either as passivating or insulating layers.

4.5.1 Wet etching

Wet etching using liquid chemical etchants is the technologically-simplest method of removing material. A wide variety of wet etches exist, for virtually every material and application conceivable, and these are well-established in III-V device fabrication.

Wet etching relies on chemistry between reactants in the liquid etchant and the surface to be etched. There are three steps to the etching process; the reactants must be delivered to the surface by diffusion, the reaction needs to occur, and the products need to be transported away from the reaction surface, again by diffusion [164]. If the transport steps are the limiting factor to etch time, then the process is said to be diffusion-limited.

If the reaction step is the limiting factor, then the reaction is said to be rate-limited. Which of these dominates is dependent on the etch composition and conditions; increased etch concentrations will increase etch rate, whilst agitation or increased temperature can affect either situation. Diffusion-limited etches are common due to the small diffusion coefficient in liquids. Wet etches are hence extremely sensitive to changes in temperature, concentration and composition, and can therefore suffer from repeatability issues.

Wet etches can be highly undamaging and can produce very smooth morphology, though many wet etches have strong anisotropic tendencies when etching semiconductors, as a result of their extreme dependence on crystal orientation [164]. As a result, the only real control of etch isotropy is substrate orientation, unlike dry etching, where control is much easier. As a consequence, when wet etching GaAs with many etchants, for example, if a rectangular mask is aligned along the dominant crystal planes, two edges will have etched sidewalls with positive gradient, the perpendicular edges a negative gradient. A 45° rotation of the mask on the substrate will result in vertical sidewalls. This can be a significant concern for deep wet etches in particular, and can considerably increase design complexity.

In addition, wet etching suffers from undercutting of the mask used to define features as result of capillary action in the fluids. Etch solutions inevitably travel under the mask and etch the underlying material, resulting in deviation from the intended feature size, with subsequent limitations in realisable density and uniformity. Though this can be minimised by using adhesion promoters or hard baking the resist masks, the problem is inherent to wet etchants. In addition, morphology of wet-etched surfaces can be highly variable, and dependent firstly on reaction rate. Slower etches tend to produce smoother surfaces, whilst faster etches can result in rapid gas production as the reaction proceeds, producing bubbles which can significantly affect the uniformity of the surface morphology [165]. Any solid or aqueous by-products of the etch process tend to also re-deposit on the etched surfaces, leading to increased defect densities and potential contamination concerns.

One advantage of wet etching is that it is generally damage-free, assuming the etch chemistry is tailored to the material system, whilst dry etched material often suffers from damage problems. In addition, since the chemistry can be manipulated with a great degree of freedom, it is possible to create selective etches for given materials. An example is the recess etch process used in HEMT fabrication, where a succinic or citric acid and hydrogen peroxide mixture etches gallium-containing layers, but not aluminium-

containing layers. As a result, the vertical etch will terminate on removal of the InGaAs cap layer, though it will continue laterally. Since only regular glassware is required for most processes, wet etching is also very cost-effective.

Single-component etches such as acids or bases will etch many materials - metals such as titanium or semiconductor oxides can generally be etched using hydrofluoric acid, whilst potassium hydroxide etches silicon. Etching III-V semiconductors is more complex as a consequence of their zinc-blende crystalline structure. III-V etchants generally operate by oxidising the semiconductor surface, then etching the oxide [164]. As a result, any etchant must contain an oxidising agent and a dissolution agent for the oxide. Though most acids will work in combination with hydrogen peroxide, a commonly used example is sulphuric acid or orthophosphoric acid with hydrogen peroxide. Adding water allows control of etch rates. The oxidation step is believed to be diffusion-limited and electrochemical; a redox reaction occurs locally at the surface when in contact with an oxidising agent. In the presence of only the peroxide, the oxide thickness will increase and reach a finite limit as the oxidising agents can no longer be effectively delivered to the surface. The acid in solution then etches the oxidised locations, removing the oxidised III-V elements. The acid alone will do nothing more than strip the native oxide from the surface.

Since the process is electrochemical, variables (such as illumination or the presence of metals) that can alter electron flow in solution can also affect etch rate either locally or across the whole wafer. As a result, uniformity of these processes may be questionable if precision is required.

4.5.2 Plasma processing

Plasma processes include a range of electron states, including flames, arcs and discharges, and are of particular interest in modifying surfaces by depositing or removing materials. The low-pressure glow discharge or "cold" plasma state, where free electron densities are high and energetic [166], is most relevant to the processing of semiconductor surfaces.

In a plasma, though the molecules, atoms and radicals are at relatively low-temperature, the electrons are highly excited, similar to the "hot" electron case described in Chapter 3 in a semiconductor under a high electric field. The very high electron energies are sufficient to overcome molecular bonding in the gas used to form the plasma, with the consequence that reactions which ordinarily would require very high temperatures can be achieved at relatively low temperatures by the use of a plasma [166]. In some types of

plasma processing, ionisation effects in the gas are used to form reactive species, whilst the complete range of possible radicals from the reactant gas species can be formed to create myriad chemical effects. These effects can then be harnessed by control of a multitude of variables; chemistries, pressure, flow rate, temperature and power. As a consequence, deposition processes are possible for a given set of conditions, whilst different radical formation will allow etching to occur.

Plasma reactor chambers are designed with these variables in mind; the wafer rests on a cathode connected to an RF generator, generally with a frequency of 13.56 MHz [167], whilst the chamber is pumped down to high vacuum. The neutral plasma is then formed above the sample using an upper anode, whilst flow rates can be controlled for the various process gases to set pressure. The plasma does not form at the electrode surface, but rather is electrostatically confined above the surface by a positively-charged "plasma sheath" which forms as a consequence of highly-mobile electrons at the edges of the plasma [168]. The sample is then exposed to the plasma. The RF voltage drop is split between the plasma sheath at each electrode in inverse proportion to the areas of the electrodes - as a consequence most of the voltage drop will be across the sheath closest to the sample in the case where the upper anode is much larger than the cathode plate [167]. Many reactors use the entire chamber wall as an anode to maximise this effect.

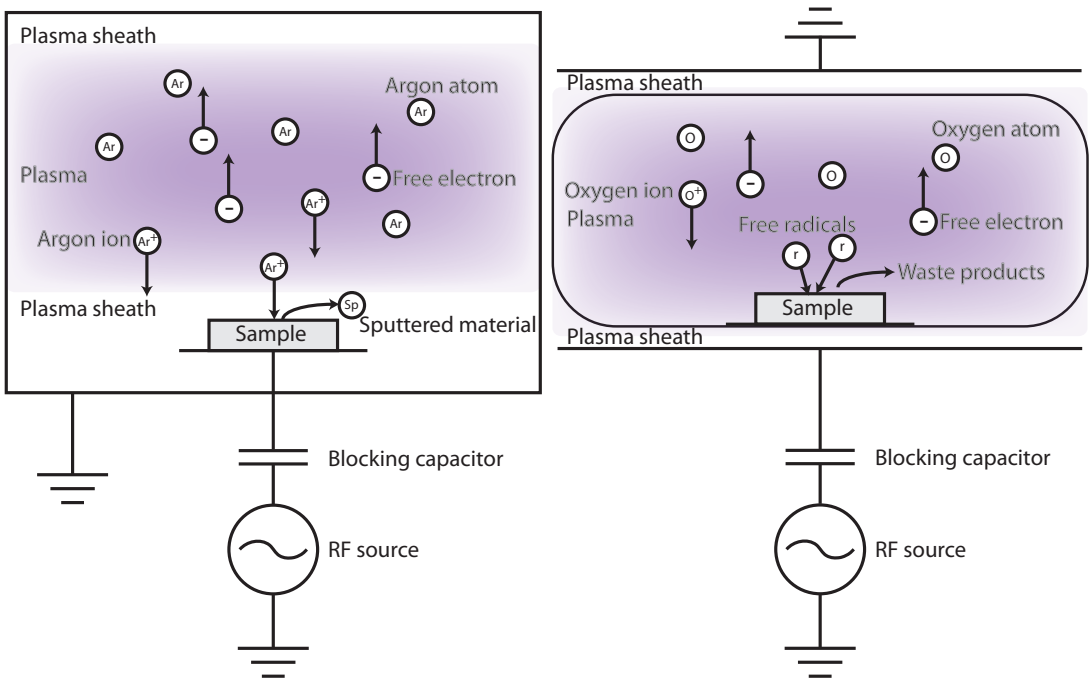
This section will describe the various applications of plasma-based processes in device fabrication.

Plasma sputtering

Sputtering is a purely physical process where the species are not reactive, and relies on momentum transfer to bombard the sample surface with energetic ions from the plasma. Inert gases such as argon are commonly used to eliminate chemical interactions with the sample. The plasma formation causes argon ionisation, and these ions are then accelerated towards the sample during the positive half-period of RF bias. On striking the surface, material is ejected from the surface. The etch rate is then dependent on the material being etched, bias and ion density. Since low pressures are usually used, the process tends to be anisotropic, with near-vertical features resulting. The process is illustrated in Figure 4.9(a).

In the context of III-V processing, physical sputtering, however, has several problems; the sputtering process generally requires relatively large energies to bombard the surface

which tend to damage the semiconductor, whilst the momentum-ejected material is not consumed by the process and is often redeposited elsewhere on the sample. As a consequence, the resultant etch profiles tend to be non-ideal and unpredictable. Sputtering is hence rarely used in controlled etch processes, but is often used to deposit films of sputtered material over a sample, as discussed in Section 4.5.4.



(a) Sputtering process.

(b) Plasma etching as used in barrel ashing process.

Figure 4.9: Overview of physical and chemical plasma processes. Particle motion is annotated during positive half-cycle.

Plasma etching

In contrast to sputtering, plasma etching relies on chemical effects between plasma-generated species and the surface to etch material. The process may be entirely chemical in nature, resulting in a generally isotropic etch profile, or may make use of the kinetic energy of the accelerated ions to enhance or enable the etch process. This kinetic assistance results in increased etch anisotropy and more directional etching. In the purely chemical case, plasma-generated radicals diffuse to the surface with randomised velocities and react. The balance of the sputtering and chemical processes will be dependent on RF power. Plasma etching processes are frequently used for resist removal, usually using

oxygen, a process also known as ashing. In the ashing case, atomic oxygen is produced by the plasma, which then reacts with the hydrocarbon resist to produce volatile (usually gaseous) products which are then removed [169]. The volatility of these products means that chemical processes tend to have low redeposition rates, resulting in a clean etched surface.

Generally, simple parallel-plate reactors are used for resist ashing, where the upper and lower electrodes are comparably-sized as additional acceleration will induce sample damage. Barrel asher configurations are also common, where the plasma is excited using electrodes outwith a sample tube, though etch uniformity is reduced. In some ashers, an earthed sheath acts to strip the highest-energy ionised species from the edge of the plasma before reaching the sample, reducing damage [170]. A barrel asher configuration is shown in Figure 4.9(b).

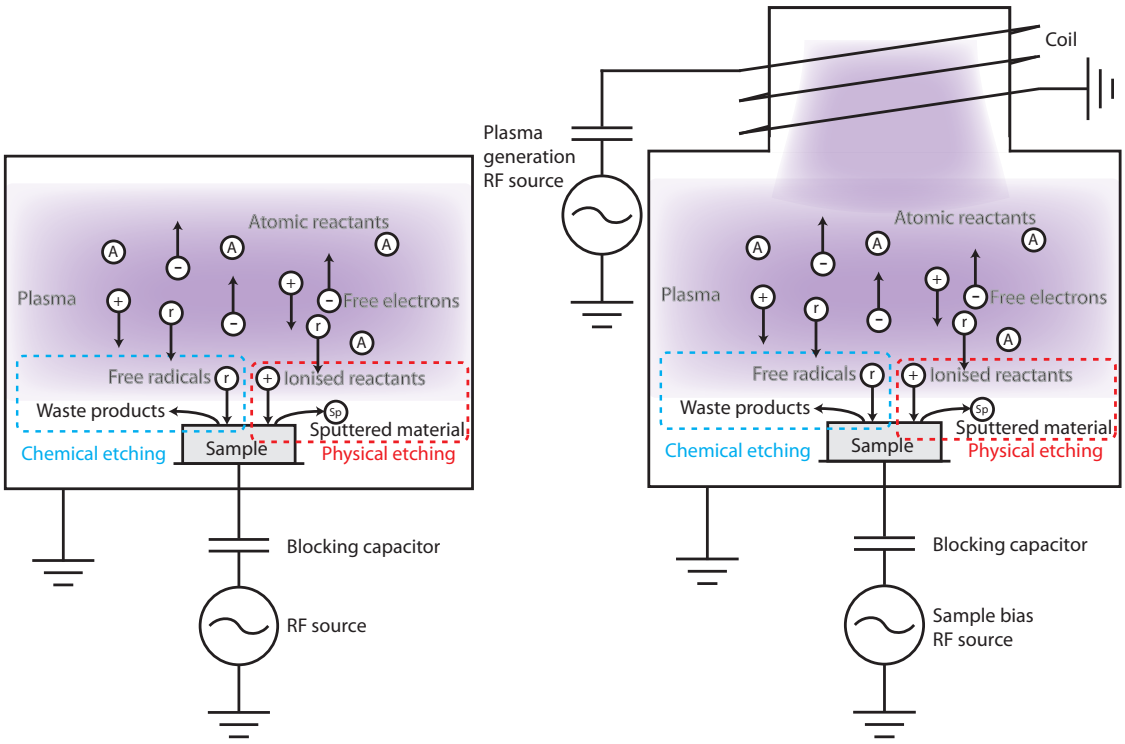
The chemical nature of the etching process also introduces an intriguing and useful possibility; selective etching. In a physical sputtering process, the only selectivity possible is determined by the relative etch resistance of the underlying material. In a chemical process, reactant products can be chosen to react with only certain materials. For example, though SF_6 will etch silicon nitride and silicon dioxide [165], it will not etch GaAs due to the thin fluoride layers that form on the surface [163]. As a consequence, layers can be etched completely by chemical processes, stopping on underlying layers of differing composition.

Reactive Ion Etching

Reactive ion etching (RIE) is essentially the same as kinetically-assisted plasma etching, but whilst plasma etching is generally used as a mostly isotropic chemical process for bulk material removal, RIE is heavily influenced by the directionality of sample bombardment by chemically-reactant ions as a consequence of acceleration across the plasma sheath, whilst retaining the benefits of etch sensitivity, selectivity and low redeposition obtained by plasma etching.

As a result, RIE reactors generally use an anode much larger than the cathode, often the entire chamber wall, as previously discussed [170]. The benefits of this are clear; enlarging the anode with respect to the cathode implies a large voltage drop across the plasma sheath adjacent to the sample surface, increasing the incident ion energy and therefore the etch anisotropy. Correspondingly, ions at the other edges of the plasma

have little energy, so there is reduced sputtering or deposition on the chamber walls.



(a) RIE process. (b) ICP process.

Figure 4.10: Overview of reactive ion etching and inductively-coupled plasma etching methods. Particle motion is annotated during positive half-cycle.

To further increase anisotropy, RIE generally uses lower pressures than plasma etching. The mean free path of the reactants is therefore long with respect to the distance from the plasma to the sample, and scattering within the reactant gas is reduced, resulting in increased directionality [171]. At lower pressures, however, convective cooling is less effective, and the samples can more easily be heated by the plasma reactions. Some RIE reactors therefore incorporate cooled electrodes to reduce the possibility of sample damage or resist flowing during etching. A further effect of the reduced pressure is a reduced etch rate, since the reactant density is reduced. As a consequence, the RF power may be increased to increase etch rate, but this occurs at the expense of additional damage. To counter this effect, some RIE systems use an inductively-coupled plasma (ICP) technique, where one source is used to control the density of the plasma, allowing high powers to be used, whilst a second source controls the plasma sheath potential, reducing the possibility of damage whilst retaining high etch rates and anisotropy.

A further benefit of RIE is that damage is greatly reduced over more physical processes; most RIE processes, though their anisotropy is heavily influenced by the ionic energy, are predominantly chemical processes, and little sample sputtering occurs: crucial for substrates such as III-V HEMTs. RIE's foremost advantage, however, is the extreme anisotropy that can be achieved; a feature than becomes critical at increasingly small dimensions as mask undercutting enlarges transferred patterns[171].

RIE processes were carried out in the course of this work using an Oxford Instruments Plasmalab System-100.

4.5.3 Dielectric plasma deposition

With suitable configuration, plasma processes may also be used for thin film deposition purposes, particularly of interest in depositing dielectrics. Thin films of dielectric have many uses in device applications; they may be used as spacer layers to separate contacts, support structures, as insulator layers for forming capacitors, or as encapsulation layers to protect devices from the environment.

Thin film dielectrics can be deposited using CVD techniques, but there are issues with their use; principally, they require elevated temperatures akin to those used for the original material growth [172]. Elevating temperatures to these levels can present problems both with damage to epitaxially-grown thin semiconductor films and with any structures subsequently formed. For example, neither Schottky nor ohmic contacts are likely to survive temperatures upwards of 500 °C. CVD is therefore rarely used in the deposition of dielectrics. Plasma-induced films can be deposited at relatively low temperatures, however, and are hence far more suitable to a device process flow. This technique is known as plasma-enhanced chemical vapour deposition (PECVD).

PECVD relies on the same fundamental principles as plasma etching; reactions which would otherwise require high temperatures to provide the necessary activation energies can be achieved at low temperatures by utilising the energy inherent in a plasma. By providing gaseous reactants and exposing the mixture to an energetic plasma, the conditions can be tailored to deposit the products on a sample surface in a controlled manner. Silicon dioxide and silicon nitride are the two dielectrics most commonly deposited in III-V fabrication, though a multitude of deposition products are possible. In the case of silicon nitride, silane (SiH_4) generally provides the silicon, whilst either gaseous nitrogen or ammonia can provide nitrogen.

Ideally, assuming a nitrogen source, the reaction would proceed as [172]:

$$3SiH_4 + 2N_2 \longrightarrow Si_3N_4 + 6H_2 \tag{4.4}$$

In reality, however, since the silane source contains hydrogen and H_2 is formed as a by-product, it is also likely to be present in the film to some degree, whilst any residue in the chamber from pump grease or moisture can result in hydrogen, carbon or oxygen contaminants in the final film, resulting in deviations from the ideal stoichiometric Si_3N_4 film, which will change the refractive index, dielectric constant, stress and permeability of the layer [172]. As a result, most PECVD films are of relatively low purity, or particularly, hydrogenated [173], though this may not present a significant problem if the desirable properties of the deposited material are as required.

A recent development in plasma deposition technology is the invention of the ICP deposition variant. As discussed, PECVD, though operating at much lower temperatures than CVD, still requires somewhat elevated temperatures. In addition, since significant energies are involved in PECVD, damage can be induced in the surface semiconductor layers during deposition. If the surface layers are very sensitive, this can be a serious problem. As for ICP etching, ICP-PECVD uses a separate coil to strike the plasma, which then has a separate RF source to control the plasma density. A second coil and source are then used to bias the sample and control the deposition rate.

In the case of silicon nitride deposition, this allows the formation of a remote high-density nitrogen plasma in the upper region of the reaction chamber, which is then reacted with the silane source to form high densities of both species which can then be deposited on the sample surface with low bias power. As a further benefit, due to the high-density plasma produced and the more efficient production of highly reactive radicals, the process temperature can be significantly reduced. As a consequence of the high-density plasma, relative concentrations of reactive elements are higher than the residual or secondary contributions, resulting in a film that is both high quality and high purity.

This work has made extensive use of room-temperature ICP-deposited silicon nitride, produced using an Oxford Instruments Plasmalab System-100.

4.5.4 Metallisation

Metal may be deposited by many different methods; the principal three being plating, evaporation and sputtering. Metals such as gold are frequently plated to form thick deposited layers for interconnects, bondpads or backside processes. Plating is generally used to produce very thick layers with large process tolerances which are not generally patterned with high resolution, but does provide a method of forming very cost-effective metallic layers.

Evaporation

Evaporation techniques rely on the heating of a metallic source to a temperature at which it vaporises. All evaporation processes can therefore be considered thermal. Evaporation processes, however, fall into two classes: resistive or electron beam (e-beam). Resistive evaporators are usually termed "thermal" as a consequence of the resistive heating of the crucible in which the metal sits. As the crucible heats up, the metal contained is heated to the point of its vaporisation.

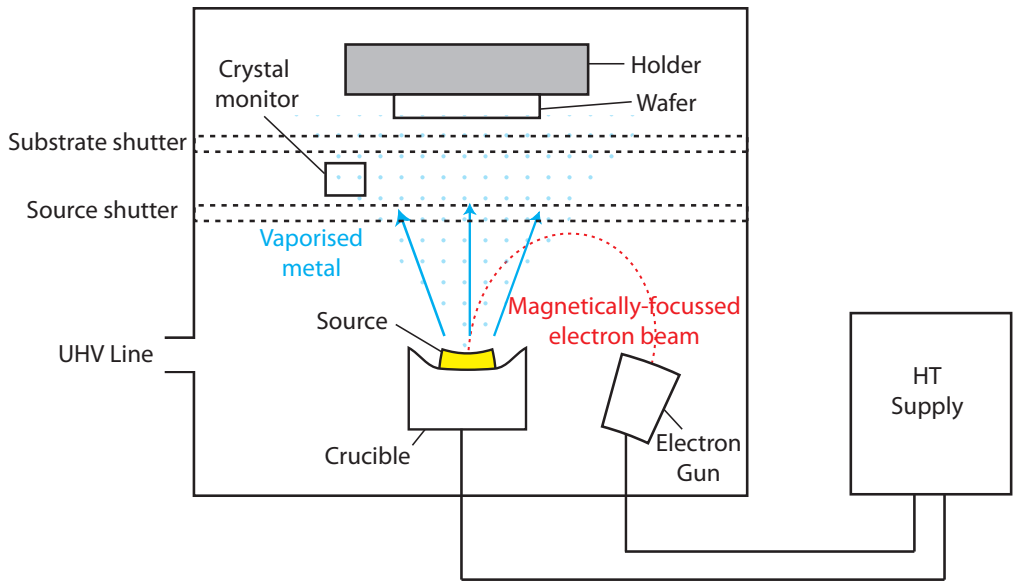


Figure 4.11: Schematic diagram of an electron-beam evaporator.

Electron beam evaporators also heat the metal, but do so locally; they do not heat the crucible and the metal sources are water-cooled in their crucibles. An electron beam

is produced using a high accelerating voltage applied between a thermionic emission filament and the crucible, which acts as the cathode. The beam can be focussed on the metal surface and as a consequence of the interactions of the accelerated electrons with the metal, it is heated locally at the focussed spot. In thermal evaporators, there is the possibility of metallic contamination by evaporating the crucible if its melting point is similar to the target metal. In e-beam evaporators, as a consequence of the local heating, there is no possibility of such contamination, yielding a high-purity deposited film. Only the locally-heated region of the metal is evaporated, whilst the rest of the water-cooled source remains solid.

There are additionally usually two shutters between the source and target. The purpose of these is to protect the sample from rate fluctuations during initial heating. When the evaporation rate is sufficiently stable, the source shutter can be opened, and the rate controlled using monitoring of the oscillation frequency of a quartz crystal and closed-loop feedback. When the desired rate is reached, the upper substrate shutter can be opened, exposing the sample to the evaporant flux. Figure 4.11 shows the salient points of evaporator operation.

All evaporation processes must take place under high vacuum in the range of $10^{-6} - 10^{-7}$ Torr, since at these pressures, the evaporating metal flux has a mean free path greater than the distance to the target. As a result, there are few collisions of the metal whilst in flux, resulting in a highly directional metal coating on the sample, holder and chamber. Consequently, evaporated films are highly non-conformal to sample topography, since the incident metal flux is usually nominally normal to the surface.

Electron-beam evaporation processes were used in this work, using Plassys MEB-450 and MEB-550 evaporators.

Sputtering

Sputtering processes differ significantly from evaporation.Whilst evaporation requires elevated temperatures for the vaporisation of the metal, sputtering makes use of physical plasma processes described in Section 4.5.2 to bombard the metallic target, and can be carried out at low temperatures. A plasma is formed using a noble gas such as argon using either a d.c. or r.f. source to excite the plasma to the activation energies required for sputtering of a given metal. The target is then negatively biased, and the plasma sputters atoms of metal from the target.

Sputtering can also be carried out at relatively high pressures around 10^{-4} Torr. As a consequence, the metal atoms tend to collide before hitting the target, randomising their angles of incidence on the sample. The resultant metallic coating is therefore far more conformal than an evaporated film. As a consequence, sputtered films cannot generally be used with the lift-off technique and are used primarily in subtractive processes. An advantage of the sputtering process over evaporation is its independence of melting point, since targets do not require to be vaporised. A wide range of metals can therefore be deposited, whereas evaporation processes are limited to materials which can be melted with relative ease.

Alloys and compounds can also be deposited in a single step, which may have particular significance in ohmic contact formation.

Sputtering processes used in this work were carried out using a Plassys MP900S sputtering tool.

4.6 Generic HEMT process flow

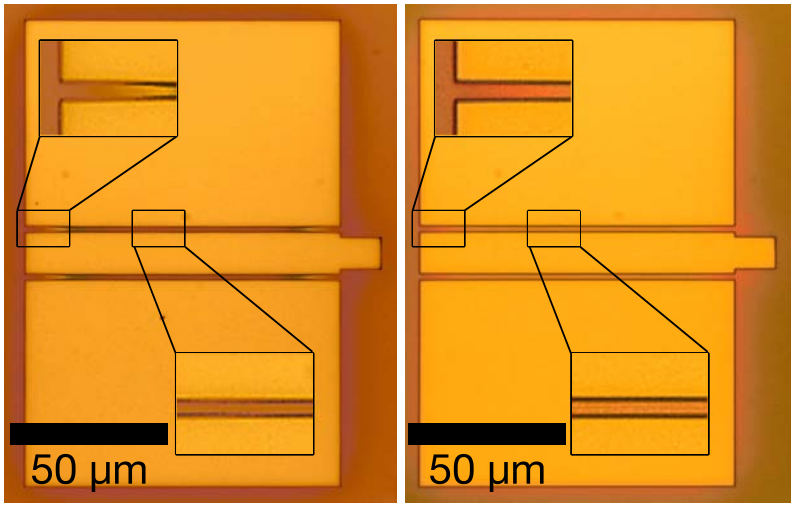
HEMT fabrication makes use of many of the techniques covered in this chapter; specifically metallisation and wet etch techniques. At Glasgow, the pattern definition is generally carried out using electron beam lithography, for its flexibility in addition to its superior pattern definition characteristics. Standard HEMT realisation combines numerous individual process steps over four key lithographic exposures. This section will aim to describe the major purposes and methods of each lithographic step in turn. The order of the processes is crucial, since more delicate structures may not survive some process steps.

4.6.1 Marker/ohmics

The ohmic contacts (Section 3.5.4) for the HEMTs are usually defined using a Au/Ni/Ge metallisation, with a total thickness of greater than 100 nm. This is achieved using the evaporation and liftoff methods described in Sections 4.5.4 and 4.4 respectively.

As briefly discussed in Section 3.1, the layout of devices is usually more complex than a simple linear design, with a multiple-finger design usual at Glasgow. As a consequence, the gate is formed between source and drain in two or more separate fingers, which are joined together with a subsequent metal deposition process. There are therefore two

source pads and one central drain pad which is shared by both gate fingers. The source-drain separation, as discussed previously, has a key effect on electric fields within the device channel, and hence must be well-controlled.



(a) No proximity correction. (b) Proximity-corrected.

Figure 4.12: Ohmic contact patterns in PMMA after development for a two-finger device.

Lithographically, therefore, three large contacts must be placed at a specified small separation in positive resist, usually on the order of 1-3 μm. This presents a lithographic challenge as a result of the proximity effect (Section 4.3.2), since regions in the middle of the contact tend to be easily overexposed, reducing the separation with the effect of variable electric field along the gate width. The required layout and the effects of proximity correction software are illustrated in the optical micrographs of Figure 4.12.

The ohmic contacts are generally annealed in most process flows, using a temperature found experimentally to produce low-resistance, morphologically acceptable contacts.

Since there are several steps of lithography required, as discussed in Section 4.3.2, markers are required to align following exposures to the first. These markers are usually metallic squares several microns in size, used in an edge location search to establish reference points across the plane of the substrate, and used to make transformations to the exposed pattern to ensure alignment. Conveniently, ohmic metallisation also gives excellent electron backscatter contrast against a III-V substrate; hence ohmic metal can also be used to produce markers for alignment in the same lithographic exposure as the contacts, reducing two exposure steps to one. One caveat is that the annealing process

must not roughen the marker edges so significantly as to cause location uncertainty.

4.6.2 Isolation

Gallium arsenide, like most III-Vs, but unlike silicon, is semi-insulating unless doped. As a result, it does conduct to a degree without doping. Adjacent devices in an amplifier setup, for example, must therefore be isolated from each other to ensure independence of operation.

This is generally achieved by either removing the material not used to form the active region of the device, or rendering it insulating by some other method. In general, a rapid, inexpensive electrical isolation process involves etching the conducting layers away; generally achieved by wet etch processes at Glasgow. The etch process must be sufficiently slow as to be well-controlled, but able to etch all the epitaxial device active layers non-selectively. Various etchants fulfil these criteria, such as sulphuric or hydrochloric acid and hydrogen peroxide. The etchant most commonly used at Glasgow, however, is an orthophosphoric acid/hydrogen peroxide mixture, chosen for its controllability and isotropic etch profiles with crystal orientation, which produce etched sidewalls with a positive gradient for any crystal plane. Concentration ratios of 1:1:100 $H_3PO_4 : H_2O_2 : H_2O$ provide repeatably controllable etch rates of around 0.7 nm/s.

The depth achieved by isolation should extend below the bottom of the channel to ensure isolation. By monitoring etched depth before and after etching, usually by Atomic Force Microscopy (AFM), this can be ensured. In addition, electrical measurements using previously-defined ohmic contacts allows a target conductance to be achieved for a given contact separation. A typical target is to achieve a current of less than 100 nA between two 150 μm pads separated by 10 μm at 2 V.

The remaining active area after isolation is referred to as the mesa. A single thick layer of PMMA is sufficient to mask the isolation etch.

4.6.3 Gates

The gate level is generally the most demanding process step to realise, due to its complex geometry and small feature size. The reasoning behind T-gate fabrication was discussed in Section 3.8, whilst the fabrication methods often used will be discussed in some depth in Chapters 6-7. In general terms, however, the gate is fabricated using a single-step process achieved using several resists of different sensitivities [159].

The gate metal is usually deposited by e-beam evaporation following a self-aligned recess etch step to strip the cap layer in the gate region. Since the aspect ratio is crucial to device operation as previously discussed, selective wet etch processes are generally used, avoiding issues with damage to the channel from plasma processing. The dimensions of the etched region are particularly important, since the recess length exposes the barrier layer directly, with a corresponding impact on the surface state density, in addition to affecting the parasitic capacitances, electric field magnitude and hence carrier dynamics and breakdown characteristics of the device.

The cap is therefore removed and the gate evaporated in a single resist step. A much larger feed structure is generally exposed simultaneously, traversing the mesa step instead of the nanometric gate. There is therefore a large additional metallic area on the mesa to ensure continuity of contact to the large bondpad structures fabricated later on isolated material.

The metal recipe used is titanium/platinum/gold, where the titanium provides the adhesion required from the small-area gate, whilst the gold provides the low-resistivity bulk of the gate. Platinum provides an interface layer to prevent diffusion of the highly-mobile gold through the titanium, which would otherwise degrade the Schottky contact to the barrier.

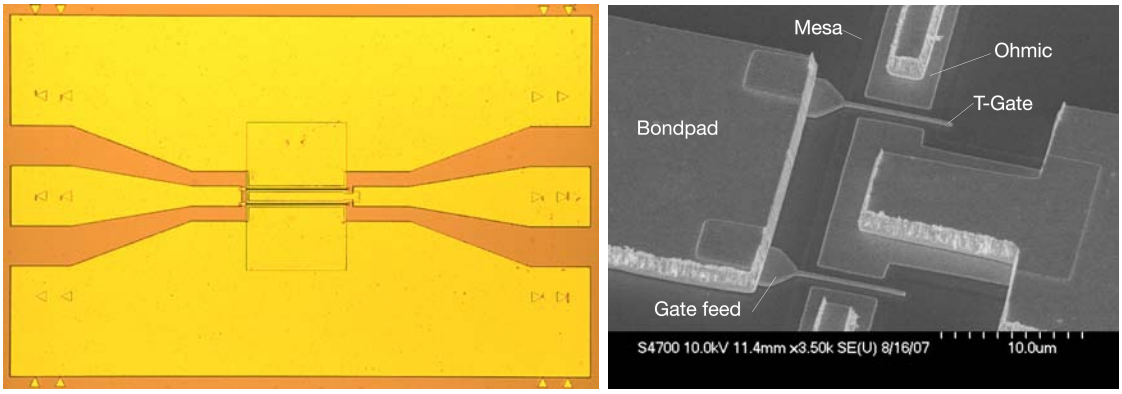
4.6.4 RF Bondpads

The bondpad layer is always defined last due to its thickness. The bondpads allow testing and measurement of the device from d.c. to r.f. frequencies.

The pads are designed as coplanar waveguides, where a ground-signal-ground topology is used to achieve r.f. signal integrity. The characteristic impedance of the line can be matched to that of the measurement system by tuning of its geometry, where the conductor widths, separations and thicknesses are variables [174]. In general terms, the conductors are much thicker than other metallised layers; up to 1.2 μm of gold is used, with a thin layer of NiCr deposited first to provide adhesion and a known resistance.

As a consequence of the metal thickness, resist around 1.5 μm thick is used to facilitate lift-off.

Bondpad definition completes a standard HEMT process flow. The general completed device layout of a standard two-finger device can be seen in Figure 4.13.



(a) Optical micrograph of completed 200 μm-wide devices after bondpad definition.

(b) S.E.M. micrograph of completed 10 μm-wide devices showing relative bondpad thickness.

Figure 4.13: Micrographs of completed standard 50 nm devices after bondpad liftoff.

4.7 Self-aligned process flow

Previous work at Glasgow was undertaken by D. Moran [175] to develop self-aligned HEMTs based on the principles of Mishra, et al., [176], where the T-gate is defined after the definition of markers and device mesa, then ohmic metal is lifted off over a blanket region over the gate. The result is source and drain contacts self-aligned to the gate, which results in reduced access resistances, defined by the gate geometry. The process is outlined in Figure 4.14.

One restriction is that the metal thickness used to define the ohmic contacts must be less than the gate foot height. In addition, since the bulk of the gate is realised in gold, there are serious thermal constraints to avoid distortion of the gate geometry.

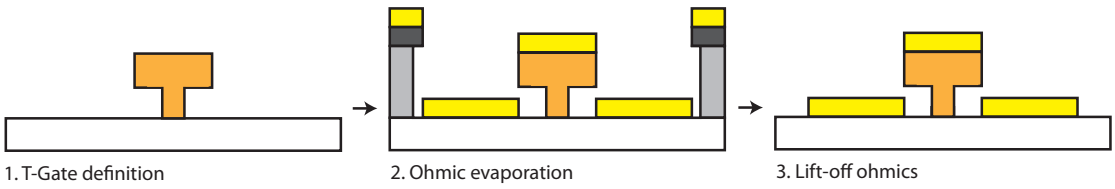


Figure 4.14: Schematic of a self-aligned process flow.

A non-annealed ohmic process was developed at Glasgow, by tailoring the energy barriers of the epitaxial structure to lower access resistances. Multiple delta-doping planes were

used to pin the conduction bands and reduce the barrier magnitudes.

As a consequence, a typical self-aligned process flow is:

1. Markers (defined separately from ohmics.)
2. Isolation
3. Gates (defined on a relatively planar substrate.)
4. Ohmics (non-annealed.)
5. Bondpads

It was also noted that a process flow incorporating the gate first yielded improved lithography as a result of the relative resist planarity surrounding the area for gate pattern definition [177].

4.8 Summary

This chapter has summarised the techniques available for the realisation of III-V devices together with the relative merits and disadvantages of each. Epitaxial growth was described, whilst metallisation, deposition and etching processes were compared, with a particular focus on the profiles and selectivity achieved and any damage incurred. Lithographic techniques were also compared, and the reasoning behind the use of electron beam lithography processes outlined. A thorough review of the fundamentals of electron beam lithography was also given, with particular emphasis on the factors that limit achievable resolution and defined feature size. The following results chapters will outline the deployment of these techniques in advanced device processing.

5. Characterisation and metrology

5.1 Introduction

Successful device development relies on the crucial processes of measurement and metrology that provide benchmarks for the performance of both epitaxial material and the devices fabricated on it. Accurate measurement provides an understanding of the underlying phenomena that dictate device performance, allowing device fabrication to be tailored to the improvement of these metrics.

Understanding of the materials structure on which the device is fabricated is therefore invaluable, since it defines ultimate performance to such a large degree. Techniques which characterise the material and contacts made to it are hence of great importance to the device engineer, in addition to device d.c. and r.f. measurements.

This chapter outlines some basic techniques used to characterise semiconductor material and device performance. In particular, the van der Pauw method for extracting carrier mobility and the Transmission Line Method for determining contact and sheet resistances are described. There is then a discussion of the process of characterising devices, both at d.c. and r.f. and the extraction of pertinent figures of merit for both.

5.2 Material characterisation and the van der Pauw technique

The quality of epitaxial material used for HEMT fabrication determines the ultimate potential performance of a device, since the electron density and mobility of the active layers define, as outlined in Section 3.6, the drain current of the device, its high-frequency and associated noise performance. In addition, effective scaling of the material is crucial for a large transconductance to be maintained as devices are scaled; a strenuous requirement whilst maintaining high mobility and electron density.

The most common method of material characterisation after growth is the Hall technique, though capacitance-voltage and current-voltage measurements from diodes are also very useful.

The Hall effect is based on the fact that charged carriers in semiconductors are deflected by both electric and magnetic fields. Magnetic fields exert a Lorentz force on carriers perpendicular to both the field and the direction of carrier motion, by a left-hand rule. Measurement of the Lorentz force therefore allows the measurement of electron (or hole) velocity of a semiconductor sample in the influence of a magnetic field. In conjunction with measurements of electron current density, the electron density for a given electric field can be extracted [178].

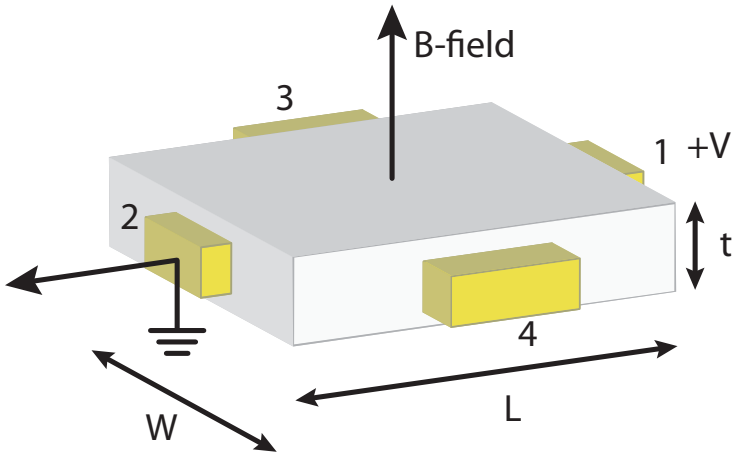


Figure 5.1: A Hall effect measurement setup.

Considering an n-type semiconductor in a uniform perpendicular magnetic field (Figure 5.1), and an applied positive voltage between contacts 1 and 2, electrons will drift from 2 to 1. The electrons experience a Lorentz force in the same plane as a consequence of the perpendicular magnetic field, deflecting them towards contact 3. Electrons accumulate at contact 3, creating a net negative charge, whilst contact 4 becomes depleted, with a net positive charge. An electric field is therefore established, opposing the Lorentz force. When this Hall electric field compensates the Lorentz force, there is zero net force on electrons perpendicular to drift and the charge transfer reaches equilibrium. This is the Hall effect.

The Lorentz force, f_L, where B is magnetic field, v_{12} drift velocity between contacts 1 and 2 and q and has its former meaning, is:

$$f_L = qv_{12}B \tag{5.1}$$

The force attributed to the Hall electric field, E_{34} is:

$$f_H = qE_{34} \tag{5.2}$$

Hence, at equilibrium, the Hall voltage measured between contacts 3 and 4, V_{34}, for a given sample width, W is:

$$V_{34} = E_{34}W = v_{12}BW \tag{5.3}$$

Current due to drift between contacts 1 and 2 is:

$$I_{12} = Wtnqv_{12} = Wtnq\mu E_{34} \tag{5.4}$$

$$v_{12} = \frac{I_{12}}{Wtnq} \tag{5.5}$$

Therefore, substituting into Equation 5.3:

$$V_{34} = \frac{BI_{12}}{tnq} = \frac{R_H BI_{12}}{t} \tag{5.6}$$

Where t is sample thickness and $R_H = \frac{1}{nq}$ is the Hall coefficient.

By measuring the applied voltage, V and corresponding drift current, I_{12} and Hall voltage, V_{34} for a sample of given dimensions, the Hall coefficient can be derived from Equation 5.6, allowing electron density to be calculated.

Conductivity, σ, is specified by the current as:

$$\sigma = \frac{I_{12}L}{V_{34}Wt} = nq\mu \tag{5.7}$$

The resistivity, ρ, is the inverse of conductivity. Mobility, μ, can also be extracted with ease from these variables, as [178] :

$$\mu = \sigma R_H = \frac{nq}{\rho} \tag{5.8}$$

The van der Pauw technique [179] is an evolution of the Hall effect, pioneered by L.J. van der Pauw. He noted that whilst Hall measurements yielded useful measurements, they depended on precise sample geometries, as evidenced by Equations 5.3 - 5.7: hard to achieve. He proposed a technique whereby the resistivity and Hall coefficient could be extracted from a sample of arbitrary geometry, with the contacts at arbitrary locations around its periphery.

In particular, if the sample was symmetric such that $\frac{V_{12}}{I_{34}} = \frac{V_{23}}{I_{14}} = R$, the situation was further simplified, leaving an equation with a single unknown:

$$\exp\left(-\pi\frac{t}{\rho}R\right) = \frac{1}{2} \tag{5.9}$$

Hence, the resistivity can be extracted with a single resistance measurement:

$$\rho = \frac{\pi t R}{\ln(2)} \tag{5.10}$$

Sheet resistance, R_{sh}, is a useful measure of two-dimensional resistivity, particularly when analysing thin films. It is simply found thus:

$$R_{sh} = \frac{\rho}{t} = \frac{\pi R}{\ln(2)} \tag{5.11}$$

Van der Pauw also showed that the Hall coefficient could be found by measuring changes to one of the previously-measured resistances, R under variable perpendicular magnetic field. The change in magnetic field drives a change in the resistance, thus:

$$R_H = \frac{t}{B}\Delta R \tag{5.12}$$

It is further noteworthy that since mobility is the product of both conductivity and Hall coefficient, it is possible to extract figures for mobility and sheet resistance without requiring a measurement sample thickness.

Considering the requirement for contacts to be placed around the periphery of a symmetrical sample, the structures shown in Figure 5.2 are commonly used for Hall measurements in HEMT structures. The first structure is a simple van der Pauw sample, with the contacts around the periphery of the measurement area attached by etched "arms" to a central square. The second has the Hall measurement area etched off using a selective etch. By removing the cap, the Hall measurements are performed on only the channel, removing the parallel conduction and giving a realistic metric for channel transport.

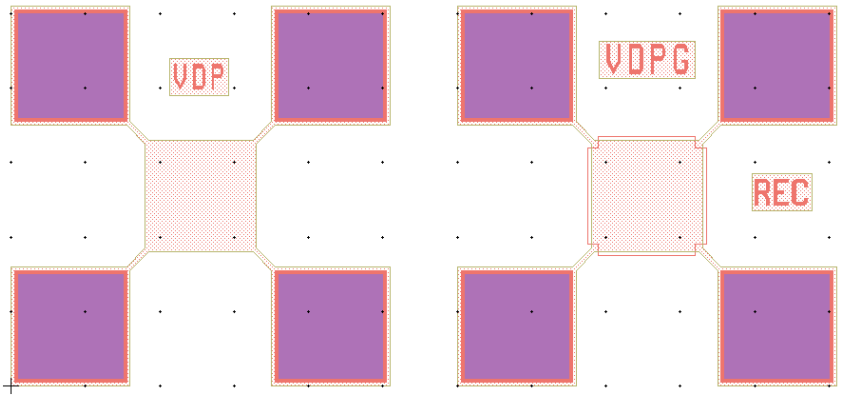


Figure 5.2: Capped and recessed van der Pauw structures.

It is essential to note that, as previously mentioned, mobility is field-dependent, with only low-field values extracted using the Hall technique. In bipolar structures, the situation is complicated by the flow of both charge carriers, whereby the Hall voltage is generated by the diffusion of both electrons and holes. In HEMTs, hole concentrations are sufficiently low below impact ionisation field thresholds that they may be neglected and the Hall / van der Pauw method assumed accurate.

5.3 Contact resistances and the Transmission Line Method

In the fabrication of devices, performance is fundamentally limited not only by the epitaxial structure of the material, but also by the ability of the device to access the potential of the material. If contacts to the material are poor and ohmic contacts are therefore high-resistance, then the device performance will be greatly limited as discussed in Section

3.7.

As a consequence, measuring and optimising contact performance is crucial to attaining the potential of devices fabricated on a given material system. In particular, a method is required to measure the resistance of the ohmic contacts.

The Transmission Line Method (TLM) is the usual technique employed to measure contact resistance. The method was first presented by Shockley, et al. [180], who proposed the measurement of the resistance of strips of semiconductor of various lengths with a constant width, with the length defined by conventionally-defined ohmic contacts. As a consequence, for a given sheet resistance, the resistance measured should be proportional to the gap length between contacts.

By extrapolating the general trend for zero length, the total contact resistance of the two contacts can be found, whilst the derivative of the resistance with length gives the sheet resistance. The transfer length, L_T describes the minimum contact length required for effective ohmic behaviour, the length over which $\frac{1}{e}$ of the total current is transferred into the material [181]. Contacts which are much shorter than the transfer length therefore have greatly increased contact resistance [182]. The process was further refined by Reeves and Harrison [183], who were able to extract the specific contact resistances with increased precision, taking alloying effects under the contacts into account. The method is shown in Figure 5.3, where R_C is contact resistance, R_{SH} sheet resistance in the bulk, R_{SK} that under the contacts and l is the contact separation. W is contact width.

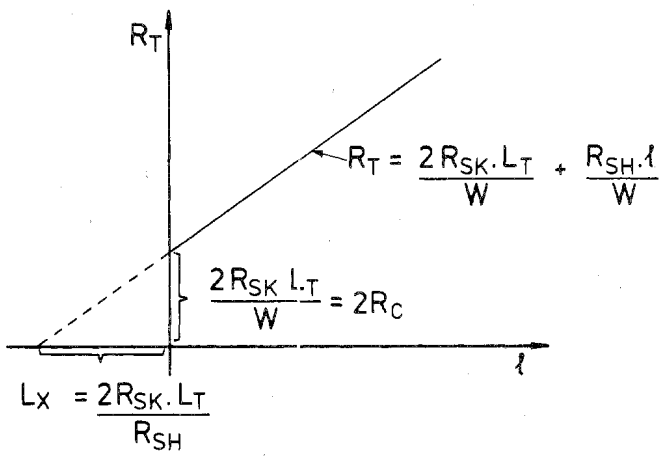
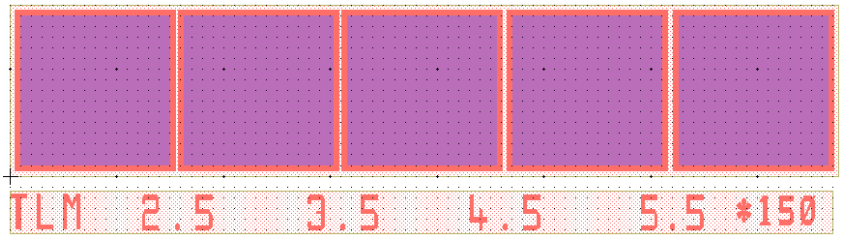


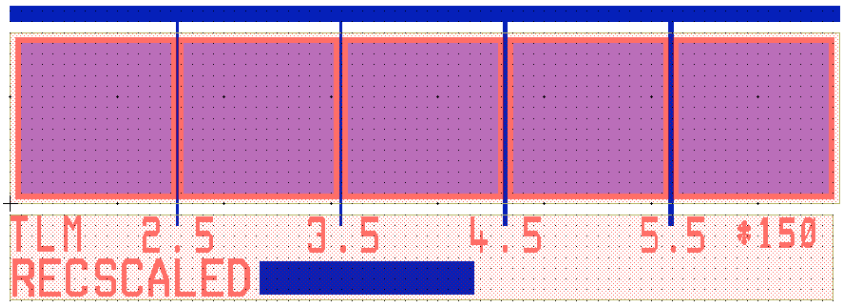
Figure 5.3: Method of extracting contact resistance, sheet resistance and transfer length from TLM resistance measurements. After Reeves and Harrison, [183].

Specific contact resistance refers to the area-related conduction properties on the specific contacts, taking current crowding (transfer length) effects into account [183]. As a consequence, whilst contact resistance has units of Ω.mm, specific contact resistance has units of Ω/mm^2. This work quotes contact resistances as Ω.mm, particularly useful when the contact length is much greater than the transfer length.

The TLM technique provides a simple method for the extraction of contact resistance, requiring simple I-V measurements of standard ohmic structures to characterise a multitude of useful parameters. The standard structures used are shown in Figure 5.4(a). The standard gap lengths are 2.5 - 5.5 μm in 1 μm steps.



(a) Capped



(b) Recessed

Figure 5.4: Capped and recessed TLM structures.

Four-probe measurements are generally used, with the voltage applied across one pair of probes and the current measured by the second pair, allowing the effect of the resistance of the probes and cables to be removed. The resistance is then extracted and the recessive best linear fit to the data used to extrapolate for contact resistance.

Measurement of the contact resistance to the cap provides useful information about contact to the cap, but as a multi-layer structure, information about the quality of the interface to the channel becomes crucial. When contacts are annealed, alloying occurs

between the deposited metal and the semiconductor surface on which it is deposited, ensuring a low-resistance contact is formed. This alloying process results in the diffusion of the elements of the alloy throughout the underlying material, reducing the resistance between the cap and channel.

In the case of non-annealed contacts, the specific contact resistance to the channel directly becomes important, as discussed in Section 3.7.1. Figure 3.17 is reproduced here for reference and references to resistances refer to this figure [58].

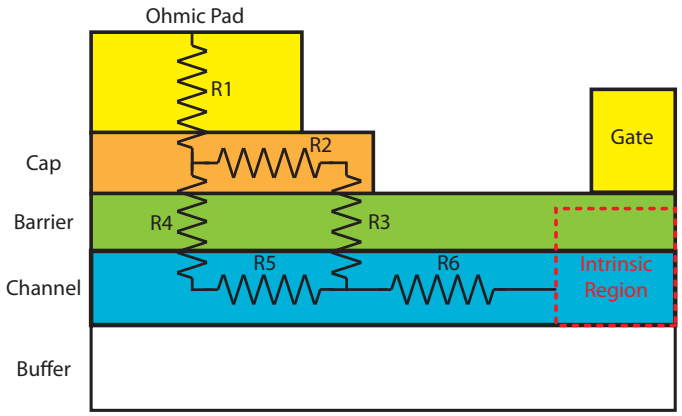


Figure 5.5: Parasitic resistances and their origins in contact and relative sheet resistances of the cap and channel.

As a consequence of the parallel conduction paths where $R_3 = R_4$, the effective total contact resistance to the channel is not entirely defined by the capped contact resistance R_1, as it would be in the annealed case; rather, the resistance is the parallel combination of $(R_2 + R_3)$ and $(R_4 + R_5)$ in series with R_1 and R_6.

Since these resistances are then largely determined by epitaxy, a method for ensuring the formation of low-resistance contacts to the channel is crucial to the realisation of high-performance devices.

The recessed TLM structure, shown in Figure 5.4(b) is one way of extracting meaningful numbers. For each gap dimension, a correspondingly-scaled region in which the cap is removed is also formed, such that the capped region remains a constant length for each, whilst the etched region varies. The recessed region is therefore effectively the TLM length, not the length between contacts directly.

A measurement of the resistance between two contacts therefore comprises twice the total

capped access resistance region shown in Figure 5.5, plus the length of the respective channel section defined by the gap length. Since the capped region is unscaling, the total resistance of the access region should remain unchanged for each. As a consequence, a linear fit to the resistance data using the previous criteria yields a contact resistance figure that includes these access regions at both contacts.

Since the dimensions of the capped region are known, and its sheet resistance is extracted using the standard capped TLM, the resistance of the capped component can be removed from the measurements [184]. The sheet resistance measured for a given capped structure will be considered to be largely due to the parallel combination of R_2 and R_5 due to the distributed nature of the resistances. The effective resistance of both capped regions can therefore be subtracted from the extracted "contact resistance" figure for the recessed case, leaving a more representative metric for the total resistance of $R_1 + R_4$. In the case of a real device, this would also be a method of extracting parasitic source and drain resistances [184].

Although this method is imprecise, and likely grossly oversimplifies the complex problem of parallel transport in the capped region for the non-annealed case, it nonetheless provides a more comparative metric for contact resistance in the recessed situation: useful for comparing material systems and fabrication processes. In effect, it allows comparison between the capped and recessed cases, where in the ideal scenario, the extracted recessed contact resistances should approach the capped case.

5.4 Device characterisation

Since devices are realised in coplanar waveguide layout as described in Section 4.6, the measurement setup must match this. As a result, both d.c. and r.f. measurements make use of three-signal probes in a ground-signal-ground configuration, at the heart of the measurement system. The probes are mounted on precision manipulator arms that allow three-dimensional positioning of the probes to attain reliable contacts.

A general system setup requires a network analyser, able to supply and analyse a range of signals of variable frequency, matched in specification to the probes and cabling. It is these components which define the measurement range of the system. A semiconductor parameter analyser (SPA) is also required to provide the d.c. measurement and bias capability. The systems are configured such that the r.f. components are disabled during d.c. measurement, with the d.c. bias injected directly to the probes. During r.f.

measurements, the d.c. bias is used to provide the desired operating conditions for the device, whose response to an input spectrum generated by the network analyser is then measured. Both systems are connected using a General Purpose Interface Bus (GPIB). Parameters can be passed and one system may control the measurement timing, whilst computer control can additionally be exerted over GPIB.

The measurement setup used in this work varied, since the systems were upgraded during the course of this project. Initial device measurements were taken using Picoprobe probes, mounted on a Karl Suss probe station using a Wiltron Vector Network Analyser (VNA) with an Agilent 4155B SPA. The system was capable of r.f. measurement up to 60 GHz.

This setup was later replaced by a semi-automatic probe station able to make measurements under automated control at a given sample location, then move to another site and repeat measurements: a far more flexible and efficient setup. The new system also uses Picoprobe probes, connected via Agilent frequency extender arms to an Agilent E8361A PNA network analyser and N2560 test set, capable of measurements up to 110 GHz.

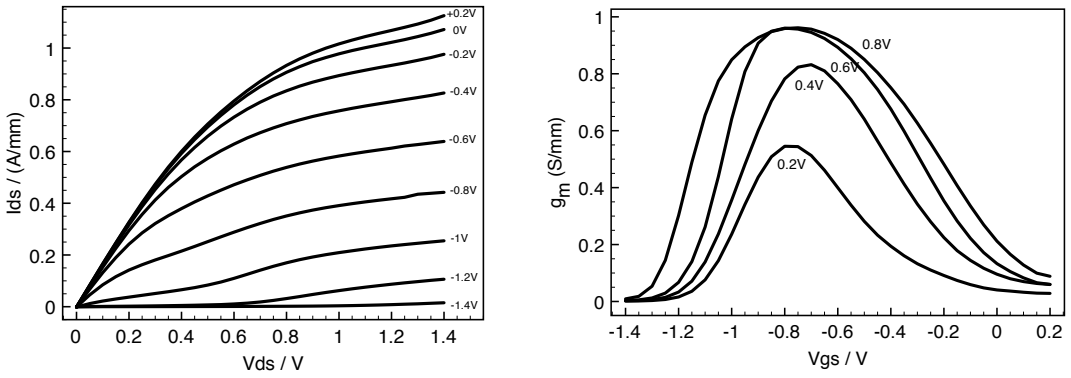
The manipulators are mounted on a Cascade Microtech Summit 12000 semi-automatic probe station, which is able to move between sample sites, but is incapable of altering probe separation. The system is connected to a control computer running Cascade Microtech Nucleus software. This computer also runs the Cascade Microtech WinCal software which allows the configuration and management of r.f. measurements. D.c. measurement capabilities are provided by an Agilent B1500A SPA, which runs Agilent EasyEXPERT measurement software. All systems are connected by GPIB, and by the end of this project the system could be configured for fully automated multiple measurements at either d.c. or r.f. for a variety of bias conditions, allowing large volumes of data to be efficiently extracted.

5.4.1 D.C. measurements

As outlined in Section 3.6.2, HEMT output characteristics depend on the application of a variable drain bias, modulated by the gate voltage.

The B1500A provides a method of measuring the drain current whilst sweeping the drain voltage and keeping the source earthed. The gate voltage can then be stepped to a new value and the drain voltage sweep repeated. By repeating this process, the complete device characteristics can be extracted. Figure 5.6(a) shows general HEMT I-V characteristics as measured by the system. Gate current resulting from leakage

through the imperfect Schottky contact is usually also measured. The system is also able to extract additional measurements from the measured data. These data can then be exported to various file formats for processing.



(a) $I_{ds}/V_{ds}, V_{gs}$ characteristics (b) Transconductance/V_{gs} characteristics

Figure 5.6: Measurements as automatically extracted for 60 nm 100 μm-wide conventional HEMTs using the B1500A.

The transconductance can be extracted using similar processes. The drain current is measured whilst the gate voltage is swept, then the drain voltage is incremented. The SPA is then used to calculate the derivative of the drain current with respect to the gate voltage: the transconductance. The profiles for a typical device are shown in Figure 5.6(b).

The I-V profile, peak drain current, gate leakage current and transconductance hence allow a great deal to be concluded about the quality of the devices, material and their processing. From these measurements, the d.c. metrics of the devices can be easily compared and extracted. All these measurements can then be repeated across many sample locations using the autoprober.

5.4.2 R.f. measurements

R.f. measurements rely on the treatment of a circuit as a multi-port device, whereby each port has a corresponding input and output signal, as discussed in the treatment of equivalent circuit models in Section 3.7.2. Single device measurements such as those taken in this work treat the HEMT as a two-port network. At r.f. frequencies, signals behave like waves on a transmission line according to Maxwell's equations, where fractions of each applied signal are transmitted or reflected, depending on the load and matching

conditions as well as the line itself. Applying an input signal at one port results in a corresponding output on both ports as a consequence of the transmission and reflection of the original signal. These transmitted and reflected signals can be characterised in magnitude and phase over a given frequency spectrum, for a given range of bias conditions. The general setup is shown in Figure 5.7.

Figure 5.7: General overview of a two-port network and its input/output signals.

Measuring the effect of the network on the input signals with respect to the output over the operating conditions hence provides a measure of the device performance.

The signals at the two ports can be characterised by a variety of methods: namely Z-, Y-, H- or S-parameters. Each in a slightly different way describes the action of the network, relating the output signals to the inputs as a matrix of parameters. A two-port network hence generates a 2×2 matrix, and so on.

Scattering parameters (S-parameters) are most commonly used, since each signal can be isolated without requiring short or open circuits, requiring only termination with a matched load to eliminate reflection in a given direction. Measurements can be easily taken in an automated fashion.

The S-parameters of a two-port network relate to the inputs and outputs thus [185]:

$$b_1 = S_{11}a_1 + S_{12}a_2 \tag{5.13}$$

$$b_2 = S_{21}a_1 + S_{22}a_2 \tag{5.14}$$

The effect of one input signal can then be eliminated by the application of a matched load, allowing each s-parameter to be uniquely described as a ratio of input and output signals

for a given load condition. In particular, S11 and S21 are extracted by terminating the output port with a matched load, whilst terminating the input port with a matched load yields S12 and S22. The S-parameters are generally plotted on a Smith chart, which, whilst in reality plots reflectance, is easily converted to impedance. By extracting all four S-parameters, the behaviour of the network over the measurement range can be characterised for a given bias point. By repeat measurements over a bias spectrum, the complete device behaviour can be characterised [186]. An equivalent circuit model as described in Section 3.7 can then be fitted to the S-parameter data.

5.4.3 Calibration

The system requires calibration in advance of its use to ensure the removal of any errors in measurement across the frequency spectrum introduced systematically by the system itself or the surroundings. Such errors would occur for all measurements, and can therefore be removed by the characterisation of structures of a known response.

Various strategies exist for calibration, such as the SOLT (short, open, line, thru), LRM (line, reflect, match) or LRRM (line, reflect, reflect, match). LRM and LRRM are similar, and require a different set of structures to the SOLT method. The main structures used for calibration are an open circuit, where the probes are usually elevated in the air above a substrate, a short circuit, where a vertical metallised line shorts the three probes together, a "thru" structure, which is essentially a short line connecting the two probes directly and a line of a given length, and a load structure, which is matched to the 50 Ω characteristic impedance of the system. The LRM/LRRM methods do not use an additional line, relying instead entirely on the modification of the reflection characteristics, whilst the SOLT method requires a known line length with a given signal delay.

The structures are provided by the use of an "Impedance Standard Substrate", provided by Cascade Microtech, which features trimmed resistors on the load structures.

Due to their generally superior performance, LRM or LRRM calibrations were used in this work.

5.4.4 S-parameter de-embedding

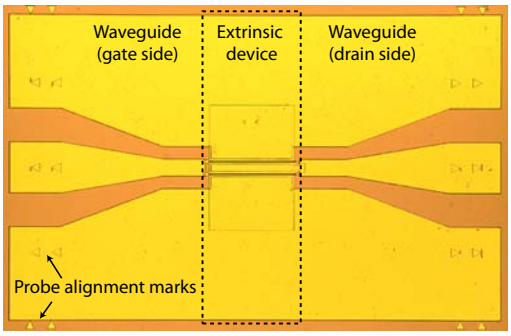
Measured S-parameters are directly useful for the characterisation of circuits and extrinsic devices. The main use of HEMTs, however, is as part of a mm-wave IC, where the

device geometry does not include the coplanar waveguides [187] required for measuring an individual device. As a result, the s-parameters measured from the devices featuring CPW bondpads, shown in Figure 5.8(a) require to have the contribution of the transmission lines removed from the measured S-parameters to more closely correspond to the extrinsic device itself [188]. This process is known as de-embedding, and it is important to note that the process removes only the contribution of the waveguides; the contribution of extrinsic device components such as the parasitic resistances should remain unaffected [189]. The de-embedding process occurs after measurement, and is additional to the essential process of system calibration.

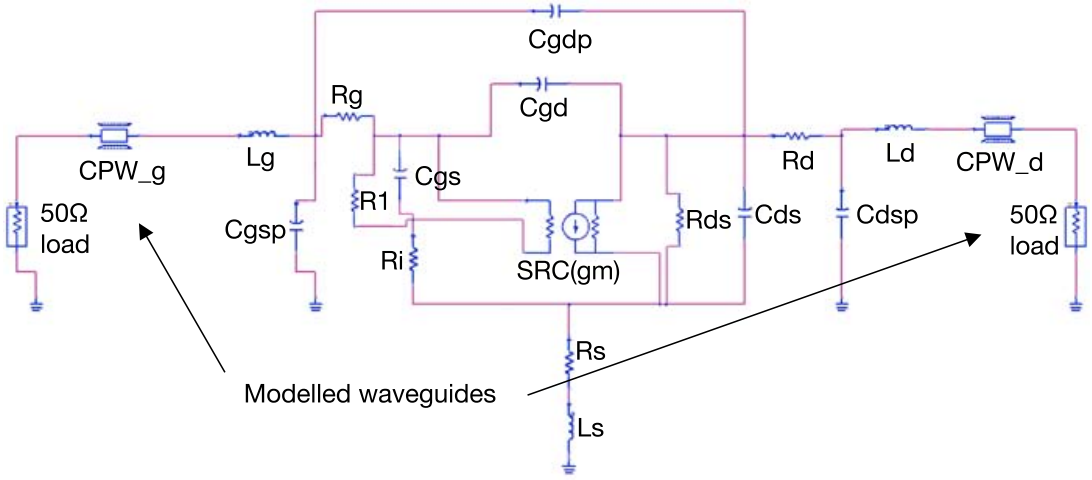
Various techniques exist for de-embedding, by identical techniques to the fitting of an equivalent circuit to the measurements . The effect of the additional waveguide is signal attenuation, and particularly a phase shift on each parameter, greatly reducing the extrinsic device performance. The line performance may be either modelled, or measured directly, then subtracted from the S-parameter measurements. A problem, however, is that the precise line length included in each device measurement is influenced by probe placement on the waveguides. Placing the probes at precisely the expected line length for each measurement is extremely challenging, especially when probes with non-ideal geometries are used or adjustments are required for optimal contact. To minimise errors and provide a guide, probe alignment structures are defined lithographically on the bondpads during device realisation. Although the use of an autoprober minimises variation between devices, human error is a major factor.

As a consequence, a useful method is the S-parameter simulation of the waveguides in software. The line parameters can then be calibrated for a real line of defined length and deposited metal thickness, defined on the device substrate, yielding an expected line length for the defined device waveguides. By then incorporating variable-length lines with this nominal line length from the calibration in the equivalent circuit, the line length can be incorporated as an additional parameter for optimisation in the equivalent model fit. The placement error can therefore be incorporated into the model on an individual device basis, whilst retaining a physical basis by calibrating the line length with a known line length.

All equivalent circuit modelling and manipulation undertaken during this project was carried out in the Agilent Advanced Design System (ADS) using the method described. The circuit is shown in Figure 5.8(b).



(a) Complete HEMT with CPW lines in place



(b) ADS circuit model for S-parameter fit, showing variable-length modelled transmission lines

Figure 5.8: HEMT transmission line modelling.

5.4.5 Extraction of figures of merit

The de-embedded S-parameters provide key information about the measured device and the device r.f. figures of merit discussed in Section 3.7.2 can be directly extracted.

The cutoff frequency, f_t, is most easily found by converting the measured S-parameters to h-parameters, extracting h_{21}, the current gain, thus [185]:

$$h_{21} = \frac{-2s_{21}}{(1 - s_{11})(1 + s_{22}) + (s_{12}s_{21})} \tag{5.15}$$

Since cutoff frequency occurs at unity current gain, by extrapolating the measured h_{21} to its intercept with the x-axis, the cutoff frequency can be found.

A further complication is that in the case of high-performance devices such as the HEMT, the figures of merit generally occur at frequencies much higher than the measurement frequency. A degree of extrapolation is therefore required. As a consequence of limited gain-bandwidth product, current gain decays at a rate of -20 dB/decade, and so can be easily extrapolated from measured results. An example is shown in Figure 5.9.

The maximum frequency of oscillation is extracted by similar techniques, though the situation is more complex. The Maximum Available Gain (MAG) is defined as [190]:

$$MAG = \frac{s_{21}}{s_{12}} \left(K + \sqrt{K^2 - 1} \right) \tag{5.16}$$

Where K, the stability factor is defined as [191]:

$$K = \frac{1 + |s_{11}s_{22} - s_{12}s_{21}|^2 - |s_{11}|^2 - |s_{22}|^2}{2\,|s_{21}s_{12}|} \tag{5.17}$$

If K is above unity, the network is unconditionally stable and cannot oscillate. If K is less than unity, the system may spuriously oscillate given certain load impedances. A device may therefore have more gain available at some frequencies than others, but may be unstable. Reliable gain in this region is known as Maximum Stable Gain (MSG) and represents the maximum gain available without oscillation at that frequency. MSG is simply defined as $\frac{|s_{21}|}{|s_{12}|}$.

MSG/MAG represents the maximum gain and may be used to extract f_{max}. MAG is also taken to decay at -20 dB/decade, hence the x-axis intercept can also be used to extrapolate for f_{max}.

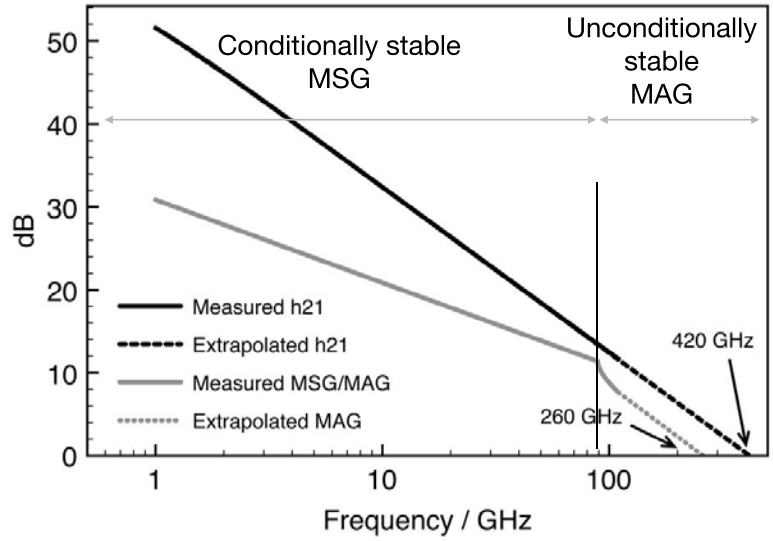


Figure 5.9: Measured and extrapolated h21 and MSG/MAG, showing extracted f_t and f_{max}.

Figure 5.9 shows the extrapolation of f_t and f_{max} from de-embedded device data from a conventional 60 nm mHEMT. The figures are 420 GHz and 260 GHz respectively. $K = 1$ at approximately 90 GHz.

5.5 Summary

This chapter has described the various techniques available for the characterisation of both epitaxial material used for devices and the devices themselves. In particular, Hall measurement systems for transport data, the Transmission Line Method for contact resistances, and d.c. I-V and r.f. S-parameter methods have been described.

The importance of correctly extracting and interpreting these data has been emphasised, in particular the requirement for calibration and accurate modelling. Since circuits based on these devices require accuracy and rely on extracted figures of merit, reliable measurement and extraction techniques are crucial. In particular, as the intrinsic device shrinks, the influence of extrinsic parasitics becomes increasingly important, and so accuracy is vital in the techniques described.

Chapters 7-9 describe the application of fabrication and characterisation techniques to the fabrication of short gate length devices within the scope of this project.

6. Literature Review

As a result of the increasing demand for low noise, high gain MMICs at ever-increasing operating frequencies, optimising the performance of the HEMT has attracted much attention. Various aspects of the operation of the HEMT have been investigated by many groups worldwide, each aiming to improve different elements of device operation.

In particular, the main areas widely investigated are the development of gate lithography, different semiconducting materials and their band engineering, the optimisation of parasitic equivalent circuit components, the analysis and engineering of surface and materials effects, and numerous fields investigating reliability.

This review shall attempt to briefly outline key progressions in the development of these.

6.1 Gate lithography

Since the operating frequency of a field-effect device is firstly dominated by its gate length, for reasons described in Section 3.8, there has been much motivation to reduce the gate dimensions as much as possible, whilst maintaining the low gate resistance required for high frequency operation (Section 3.7). As a result, T-gate strategies have generally been used for HEMTs, combining a short gate length with a large cross-sectional area to reduce resistance.

Although commercial HEMTs are commonly fabricated by optical stepper, most research work on gates has concentrated on the use of electron beam lithography, due to the small feature sizes that are achievable with relative ease, the possibilities of single-step T-gates afforded by high-voltage operation and design flexibility. As a result, as the technological capabilities of lithography tools have improved, so too have the resulting gate features and their subsequent devices.

The first T-gates were developed for MESFET applications in the late 1970s, following the work of Wolf [81] and others in determining the role of circuit elements in device performance, as a means of reducing the gate resistance of a short footprint gate. At the time, no lift-off procedure existed capable of fabricating a three-dimensional structure where the upper level had a larger dimension than the supporting structure, and as a result, various groups used photolithography, layered metals and selective etching of these in order to produce the first T-gates. This technique was first reported by Takahashi et al in 1976 [192], who used a molybdenum / gold bilayer and selective Mo dry etching to produce gates with a half-micron footprint, and a larger upper head. The technique was later refined [193] to produce T-gates with footprints as small as 100 nm, retaining a 1 μm gate head. The authors claimed no appreciable damage was caused to the underlying GaAs channel by the gate process. The technique was later used by different groups [194] for different metal compositions to produce similar results. The technique, however, relied on the precise control of the etch chemistry to ensure the exact lateral etch rate required for a given gate length was achieved.

An easier method was pioneered in 1980 by Todokoro [195], who developed some of the first resist bilayers for sub-micron electron beam lithography. The use of bilayers for improved metal lift-off had been recently proposed, using a less sensitive resist on top of a more sensitive one. Todokoro realised that by creating a contrast in solubility rates between two resists, the more sensitive resist atop a less sensitive one, using careful development, three dimensional structures of various types could be created using only a single exposure. By Monte-Carlo simulation and the use of a PMMA/MPR bilayer, the first lift-off T-gate structures were created using single-step lithography. The resultant resist profiles were as small as 200 nm with a 700 nm head, whilst lifted-off structures with footprints of 400 nm were created.

This technique was later refined by Matsumura, et al. [196] using overlay techniques to expose the more sensitive upper resist with a low dose, and a higher dose to expose only the central footprint, giving structures as small as 200 nm. This method unlocked the potential of the lift-off T-gate, allowing short gates with very large overhangs (and thus very low resistance) to be fabricated. In effect, these two contributions from Todokoro and Matsumura formed the foundations for all single-step gate lithography that was to follow.

In particular, efforts by Chao et al [197] led to the development of single-step T-gate processes realised in PMMA/P(MMA-MAA) copolymer bilayer and trilayer resist systems.

This proved to be a very simple and manufacturable process, since the resists did not intermix, and many groups subsequently made use of similar processes in the development of increasingly short gate length III-V FETs. Such a system was, for example, used at Glasgow for the fabrication of devices with gates as small as 70 nm, and is used by many groups worldwide, especially for large-scale MMIC fabrication at moderate gate lengths [198–200].

The problem with the copolymer process proved to be its low sensitivity contrast between the two resists. As a result, new resist systems were developed using more modern, faster ebeam resists, taking advantage of the advent of chemically-amplified resists. These were more sensitive than previous resists, and allowed a large contrast between the chemically-amplified resist used for the head and a conventional resist used for the foot. Various groups [159, 201] worked on systems using UVIII or UV113 for the head and PMMA or ZEP520 as the foot resist, which resulted in gates with very high aspect ratios and a fast throughput. Very stable, high yield gates, particularly at a 50 nm gate length, have resulted from this type of process; accordingly, they have been successfully employed in MMIC fabrication [202]. Although shorter gate lengths have been achieved using this process [203], these gates have never been successfully incorporated into an active device process flow, due to the intrinsic mechanical instabilities of such structures.

The drive for shorter gate lengths has therefore required more exotic fabrication techniques, as these simpler lithographic processes have resulted in unstable gates or resist profiles that simply cannot be transferred to metallised structures.

A natural evolution of the PMMA/LOR/UVIII system is the use of a two-step "bi-lithography" process, aligning the foot to the head using the same single resist stack in two exposures [204]. This avoids any effect of the head exposure on the foot PMMA, but requires precise alignment capability. Despite the lithographic improvements, the sub-30 nm foot resist profiles generated are non-vertical and have not resulted in lifted-off gates. This author has achieved similar results in Section 7.3.

The main problems associated with a single-step process involve issues of mechanical stability of the metallised gate, resist flowing and closing of the narrow gap in the resist as a result of lateral metallisation, discussed in Section 7.3.2. To combat this, a process was devised by Chen, et al. [205] to use a thin dielectric layer as a mechanical support for the gate and a means of defining the gate foot, stopping the resist from closing during evaporation. This allowed metallised gates as small as 30 nm to be fabricated, but was somewhat damaging to the underlying substrate.

The diametrically opposite view to the fabrication of T-gates is to separate the foot definition from the head entirely; a standpoint enabled by the evolution of lithography tools and their corresponding increasing capacity for accurate lithographic alignment. The first two-step method was proposed by Suemitsu, et al. [206], who used a bilayer of silicon dioxide and silicon nitride to support the gate, etched using a fullerene-enriched ZEP520 resist layer following from the work of Ishii in nanocomposite resists [207], who claimed to have achieved increased resist contrast and etch resistance by the incorporation of fullerenes. This method used the dielectric bilayer as a recess etch mask, then sputtered tungsten silicide (WSiN) into the etched trench before photolithographically aligning a larger head feature to the first, and etching away the remnant tungsten by RIE. This method produced lattice-matched 30 nm devices with cutoff frequencies of 350 GHz. Suemitsu then went on to research two-step recessing [106] using an InP etch stopper layer, which was claimed to reduce gate length extension, the kink effect and the influence of surface states in general.

This process was further developed by Yamashita, et al. [208] of Fujitsu, Japan, who used non-fullerene ZEP to etch a silicon dioxide support and an aligned ZEP/PMGI/ZEP trilayer to form T-gates with dielectric support. This method resulted in very delicate 25 nm gates, but produced devices that retained the record cutoff frequency for five years: a pseudomorphic device that achieved 562 GHz.

A two-step process based loosely around Suemitsu's approach was also reported in 2006 by Kwang-Seok, et al. [209], using a dielectric redeposition and etch technique to produce features smaller than was ordinarily possible by regular lithography. By thermally reflowing resist after patterning, or by the conformal deposition of multiple dielectrics and anisotropic RIE and then sputtering WSiN, 30 nm HEMTs with a cutoff frequency of 425 GHz were achieved.

This group later reported an evolution of their reflowing technique, using a plasma-assisted polymer deposition technique [210]. By controlling the conditions of a CH_4/H_2 plasma, a polymer may be deposited which is resistant to SF_6 etching. By then adjusting the plasma conditions, the same gas can be used to etch SiO_2. Using this technique, the group was able to fabricate sub-30 nm features, which then had a larger ZEP/UV5 layer aligned. Process uniformity, however, was highly variable, with 20-30 nm features resulting from a 25 nm desired feature size. Devices with a cutoff frequency of 450 GHz resulted.

In September 2007, this author presented a gate module for robust HEMT fabrication at

the MNE conference in Copenhagen, reporting the fabrication of robust 22 nm T-gates incorporating a silicon nitride layer to entirely encapsulate the gate recess, as detailed in Section 7.4. At the time of presentation, this was the world's shortest gate, with the expectation that the process could be scaled further. Although functional devices had been fabricated, the layer structures were unoptimised and the characteristics were poor. At the MNE conference in Athens in September 2008, improvements to this process were reported that yielded structures as small as 10 nm, by processes detailed in Section 8.3.

In December 2007, Yeon, et al. reported an evolution of their process, in which the gate was defined following the gate recess, stopping the lateral spread of metal during evaporation. This resulted in 15 nm gates, and HEMTs with a record extrapolated cutoff frequency of 610 GHz [87], though the publication contains no details of uniformity.

6.2 Materials Advances

The potential performance of a HEMT is fundamentally determined by the epitaxial layer structure, and thus the fabrication of devices can be seen as a method of accessing the transport properties of the underlying semiconductor. The performance of the devices is therefore a combination of the transport characteristics of the epitaxy, the optimising of the access regions therein and the appropriate mutual scaling of the epitaxial and device geometries.

Much attention has been given in the last two decades to the adaptation of HEMT heterostructures to epitaxial layers exhibiting higher mobility, such that the 2DEG forms in a high-mobility material. In general terms, this has resulted in the move towards the incorporation of device channels with increasing indium compositional fractions, resulting in an evolution of GaAs/AlGaAs heterostructures towards the use of InGaAs/InAlAs ternaries. As the indium fraction increases, the bandgap of the material decreases towards that of InAs, resulting in increasing mobility. Accompanying this, however, is a gradual increase in the lattice constant of the channel layer, resulting in a large mismatch to the underlying substrate and to the adjacent epitaxial layers. Gallium arsenide substrates have historically been employed, but shifts towards higher compositional fractions with larger lattice constants have led to the increased adoption of indium phosphide substrates, which themselves feature a slightly larger lattice constant than GaAs.

For many years, epitaxial growth technologies proved to generate serious lattice dislocations in non-lattice-matched situations which acted as trapping centres, and as a result,

lattice-matched HEMTs exhibited the most desirable device characteristics for many years. As growth techniques, particularly MBE, matured, however, this ceased to be the case, and the successful incorporation of strain in the device channel led to the growing performance and popularity of the pseudomorphic channel system. The subject of strain is addressed in greater depth in Section 3.4.

This fundamental shift has led to the need to engineer both substrates and epitaxy to accommodate the epitaxial strain induced by the incorporation of increasingly mismatched layer structures. As previously described, the solutions generally involve the careful matching of channel strain in pseudomorphic devices (pHEMTs), lattice matched solutions (lm-HEMTs), the use of compositionally-graded buffers to accommodate lattice mismatch in metamorphic devices (mHEMTs), or some combination of metamorphic and pseudomorphic devices, where metamorphic buffers can be used to provide a "virtual substrate" of widely varying lattice constants as the buffer layer for further epitaxy, allowing the engineering of the channel strain.

As a result of the favourable transport properties of strained InGaAs [40], pseudomorphic devices on indium phosphide have generally exhibited superior high frequency, low noise performance in the last two decades, with the majority of the top-performing devices utilising an InGaAs/InAlAs epitaxial structure on InP. In particular, the 562 GHz Fujitsu device [208] represented the state of the art for many years, and was based around a benchmark scaled layer structure on InP.

In 2007, however, this record f_t was surpassed [87], and the new 610 GHz device was based on a metamorphic material system, verifying the maturity of metamorphic growth techniques. This device used a 75% indium channel grown pseudomorphically on a metamorphic buffer.

Indium phosphide-based devices remained important, however, with ongoing research eventually yielding increased f_t and f_{max}.

The f_t figure of Yeon, et al. was superseded by Kim and del Alamo, who reported a 30 nm InAs pHEMT with a cutoff frequency of 628 GHz [12] in summer 2008, emphasising the requirement for optimisation of the epitaxial structure in conjunction with gate length reduction. This device, similarly to the earlier devices by Yamashita, et al., used a silicon dioxide support structure to provide stability for the gate and increase the foot height. Crucially, the device performance was attributed to the use of a 10 nm $In_{0.53}Ga_{0.47}As/InAs/In_{0.53}Ga_{0.47}As$ channel, with significant improvements to mobility.

A 4 nm barrier layer thickness was also used, yielding very high aspect ratios. Improvements to the scaling of the gate recess were also noted. Each of these is key to overall device performance, and major developments in each will be covered in this section. Interestingly, perhaps the key finding of the work was that the best combination of f_t and f_{max} was achieved by a 50 nm device, not the shorter 30 nm device, suggesting either the need for further material scaling or processing problems at very short gate lengths.

Although f_t is often regarded as the main figure of merit for r.f. field effect devices, it is important to realise that the maximum frequency of oscillation, f_{max} is also crucially important, since power gain may be required without corresponding current gain for many applications. In these terms, devices with f_{max} exceeding 600 GHz were reported in the mid-1990s [211] and never exceeded by field effect devices until 2007, when Lai et al., of Northrop Grumman, California [212] reported 35 nm InP pHEMTs featuring extrapolated f_{max} in excess of 1.1 THz. These devices featured a composite InGaAs/InAs channel of unreported composition. The performance of these devices was confirmed by their incorporation into MMIC amplifiers demonstrating 18 dB gain at 300 GHz. Whilst f_t impacts linearly on f_{max}, there are many other factors involved in the optimising of the maximum oscillation frequency, as described in Section 3.7.2. As a result, devices exhibiting a large f_{max} often exhibit an unimpressive f_t. In the case of the Grumman device, an f_t of 385GHz was measured; far from industry-leading for a 35 nm device.

6.2.1 Development of Metamorphic Growth Technology

Much effort has been expended on the development of metamorphic growth of HEMTs since the late 1980s [213, 214], mainly for the purposes of commercialising devices with high-indium content channels. Although it was moderately routine to realise high-quality high-indium InAlAs/InGaAs HEMT structures on indium phosphide, wafers are extremely expensive, fragile and difficult to process, and, perhaps crucially for mass market, are only currently generally available up to 4" wafer format. Gallium arsenide affords greater mechanical reliability, lower cost, less fragility and the possibility of using larger wafers [215].

In addition, metamorphic growth opens up new device possibilities since using purely lattice-matched or pseudomorphic techniques on either GaAs or InP restricts the indium content of the active layers to prevent extraneous unsatisfactory lattice mismatch. Pseudomorphic GaAs devices have generally been limited to 0-25% indium, whilst InP substrates generally preclude the use of indium compositions less than 50% [215]. By

creating buffer layers of various lattice constants, this restriction can be removed, and virtually any indium concentration can be applied to the channel. This gives a great degree of freedom in device realisation, and allows the optimising of high frequency gain with the corresponding drop in breakdown voltage that accompanies an increasing indium compositional fraction. In this way, device active layers can be tailored for their given application.

In principle, metamorphic growth should allow devices of equal electronic performance to be grown on GaAs as on InP; however, much of the work on metamorphic buffer growth over the last two decades has focussed on the elimination of lattice dislocations caused by the growth of materials of mismatched lattice constants. Although engineered strain is often desirable as a result of increased conduction band discontinuities and electron velocities and decreased scattering [40], the strain must be managed to prevent lattice dislocations which induce scattering. Early devices, though operational, were of limited performance when compared to similar lattice-matched devices, as a result of strain-induced defects [214]. As a result, most work in metamorphic device technology has focussed on the development of strain relief buffers, and many academic and commercial research groups worldwide now focus on metamorphic technologies.

Reducing lattice defect density must be done without sacrificing surface morphology, such that modern lithography is not precluded, and r.m.s. surface roughness is a reasonable measure. InGaAs [216], InGaP [217], InAlGaAs [218], AlGaAsSb [219], and various other ternaries and quaternaries [198] have been investigated, in addition to the growth techniques employed in their use, whether featuring stepwise or linearly-graded compositions [220].

Most work [87, 221–225] has used slowly graded compositional fractions leading to a strain-relaxed (lattice-matched) $In_{0.48}Al_{0.52}As$ final buffer layer for channel growth. Various studies have also concluded that metamorphic buffers appear not to be a huge concern as regards reliability, in contrast to issues, for example with gate sinking and hot carrier effects [226–228].

Before the publication of the 15 nm 610 GHz device by Yeon et al. of Seoul University, the highest f_t exhibited by a metamorphic device was 440 GHz [224], achieved at the University of Glasgow. This device featured a 50 nm gate, suggesting that there may be performance-limiting factors in the case of the Seoul device. Nevertheless, it is clear that metamorphic devices have reached a level of maturity at which they may compete with lattice-matched buffer solutions on InP for certain applications.

6.2.2 Optimisation of the Recess Region

In addition to the scaling of the HEMT's gate length, it is crucially important to maintain a favourably large aspect ratio of the gate length to the gate-to-channel distance. The aspect ratio, as for MOSFETs, is required to be sufficiently large as to allow the transconductance of the device to remain high as the gate length is scaled laterally to maintain effective channel control [229].

This can been achieved in two main ways: either by optimising the gate-channel distance in the wafer epitaxy, or by using the gate recess to etch the device barrier layer down to the desired dimensions. A combination of the two approaches is often used.

In addition to maintaining high transconductance, recess depth has also been shown to significantly affect output conductance, gate capacitance and noise performance [230]. Much of this, however, is likely affected by the interplay of electron populations in the various device layers resulting in the formation of parasitic conduction channels with varying recess depths. As a result, there may be fewer geometrically-related results from this work than publications may suggest.

As well as controlling the vertical recess depth, much research effort has been expended on investigating the effects of scaling the recess laterally with respect to the gate. It has generally been found that shorter recesses produce superior results in terms of high-frequency performance, with generally higher transconductances [231] and in particular, less profound kink in the I-V characteristics [232]. As a result, much work has been done into the investigation and minimisation of the kink effect [75, 76], resulting in its explanation in terms of surface state density in laterally-etched recess regions, amongst other phenomena. In particular, for obvious reasons of electric field, an asymmetric recess with a short gate-drain recess should theoretically yield optimal results in terms of electron transport, as has been experimentally verified by various groups [87, 233]. A more detailed discussion of the effects of recess dimensions on device performance can be found in Sections 3.6.2, 3.7.1 and 3.8.

Various methods have been adopted to achieve favourable recess profiles, split, as for all etching, into wet and dry selective and non-selective processes. Wet etch recess processes continue to be the norm in HEMT fabrication as a result of the damage generally induced by the dry etching of III-V materials, though various dry-etched recess processes have also been developed. In terms of wet etching, there exist many selective etch processes, which are most frequently used in modern process flows due to the ability to remove only the

cap layer without attacking the underlying barrier by the use of an aluminium-selective process. These are generally based around the use of pH-balanced succinic [234, 235] or citric [236, 237] acids.

There are also non-selective etch processes, however, which have made use of the "digital etch" methodology [238]. Digital etches split the oxidation-etching nature of III-V chemical wet etch processes to oxidise a finite thickness and then remove it, allowing, in principle, an extremely controllable etch depth. Processes have also been developed to minimise the lateral etch dimension with such methodologies [239].

6.2.3 Alternative Channel Materials and Designs

Various groups have recently expended considerable research effort into the development of composite channel devices, notably the devices from Kim, et al. [12] and Lai, et al., [212] mentioned previously, which achieved record f_t and f_{max} figures respectively. In general, there can be various goals in designing composite channels, but all relate to the combination of the benefits of various different channel materials, whilst minimising any drawbacks.

For some groups, the motivation is to develop devices which perform like an InGaAs-channel device at low bias conditions, but more like a power transistor at high bias conditions, by forcing "hot" electrons into a wider-bandgap material, such as InP [240–242]. This has the effect of reducing impact ionisation at high fields, making use of InP's large high-field mobility, and preventing effects such as real-space transfer into the buffer. Similar effects have been seen using InGaP as the second channel material [243].

Some work [244] has also been done into the use of multiple channels, each with their own spacers. The motivation has been to allow high electron density and high mobility, with low gate leakage and output conductance, but the work has not resulted in particularly noteworthy performance.

The most common motivation is to further enhance the channel mobility of InGaAs/InAlAs HEMTs by the use of an InAs/InGaAs composite channel. This approach follows similar methodology to the InGaAs/InP structure, but using the InGaAs as the wider-bandgap channel material [245], which acts to reduce the channel impact ionisation. A slightly different approach is to "sandwich" the InAs channel by InGaAs [231, 246], which was found to reduce strain as well as improve channel transport. In the case of both the current record devices from Kim, et al. [12] and Lai, et al., [212], an InAs sub-channel

was used between thin layers of $In_{0.53}Ga_{0.47}As$ to improve the transport properties.

Although the GaAs/AlGaAs and InGaAs/InAlAs material systems are now fairly well-established, there has, in recent years, been a surge in the development of various other HEMT materials systems, particularly in the field of high-power devices, where nitrides are now becoming dominant. As a result, there has been a great deal of recent development into GaN/AlGaN HEMTs, with a corresponding research effort into each element of the materials system. Although these high-power devices are generally only moderately high-frequency, there is much less trade-off between the two than for other materials systems.

In terms of high-frequency performance, antimonide-based material systems have recently begun to be developed more seriously, with a particular bias towards high-frequency, low power devices well-suited for mobile RF or CMOS applications. InAs/AlSb/AlGaSb and InAs/InAlSb/AlGaSb structures have been fabricated [247–249], but the real goal is moving towards the use of InSb as a channel material, which exhibits the highest mobility and saturation velocity of any known semiconductor, up to 70000 $cm^2V^{-1}s^{-1}$ at room temperature [250]. Functional InSb/AlInSb devices were reported [251–253] by Qinetiq and Intel, and, though currently designed with digital applications in mind, are beginning to report high-frequency performance comparable to the much more mature InGaAs/InAlAs devices. With increasing technology maturity, it is to be expected that InSb-based devices should eventually surpass the high-frequency performance of other materials systems. Interestingly, recent work on compressively-strained InSb [254] has yielded acceptable p-channel transistors as a consequence of its reasonably high hole mobility, a promising result for InSb logic applications given the usual absence of a corresponding p-channel device for the predominantly n-channel III-V devices designed for CMOS.

6.3 Optimising Device Parasitics

Although not the primary driver in device performance, the parasitic elements of the HEMT contribute greatly to its overall performance as discussed in some depth in Section 3.7. As a result, a reasonable degree of work has also been undertaken to reduce the effects of these parasitic elements.

Most work has been done on reducing the access resistances, since both source and drain resistances impact on high-frequency performance, with a particular emphasis on the

source resistance, which also acts to reduce extrinsic transconductance. Engineering wafer epitaxy for low-resistance channel and cap layers, as well as optimising ohmic contacts for minimum possible contact resistances results in the decrease of these parasitic elements, but beyond this, there are more fundamental ways of reducing these resistances.

The most obvious way is by reducing the access region dimensions, bringing the ohmic contacts closer to the intrinsic device region. Until very recently, due to fundamental limitations of lithography, the primary method of achieving this was to adopt a self-aligned strategy where the ohmic contacts can be defined using the gate lithography as a mask, or vice versa.

There are three main self-aligned process flows which have been adopted to achieve lower access resistances. The first method involves the use of a sidewall spacer process after gate definition, whereby an insulating dielectric is deposited over the gate, then subsequently anisotropically etched and ohmic contacts deposited [176]. This method can be used for any type of gate, though is more usual for pyramid gates [130, 255, 256] , but requires many process steps, may not result in a particularly close gate-ohmic spacing for T-gates, and can affect parasitic capacitances.

The second is the use of the T-gate profile as a shadow mask for the deposition of ohmic metal, thereby defining the ohmic separation by the length of the gate head. This has the advantage of simplicity, but is restricted in several ways. The ohmic metal must be no thicker than the height of the gate foot, the contacts can only be brought as close as the gate head length, implying an inherent performance trade-off, and there is a very limited thermal budget following gate definition to prevent deformation of the gate. Such strategies have been used very successfully [257, 258] to enhance device performance using a non-annealed ohmic contact method, and this technique has been the most common method of self-aligning high-frequency HEMTs.

The third method is to reverse the process flow, depositing ohmic metal over the complete wafer before gate lithography [13]. Gate patterning into a dielectric and subsequent recess etching can then be used to separate the contacts and achieve a close spacing, as defined by the lateral recessing. This method has resulted in relatively small gate-ohmic separations, but is much more involved, limits the techniques available for advanced gate lithography and is additionally limited by the recess process latitudes, which play a large part in defining the electric fields in a device.

Beyond minimising the access resistances, little specific work has been done on optimising

device parasitics. Various process flows acknowledge the effect of processing on parasitic capacitances [199, 245, 259] and gate resistance but whilst theoretical analysis underlines the importance of complete parasitic optimisation [4, 81], few groups have focussed on reducing these. A limited amount of work has been done on the modelling of parasitic capacitances and gate resistance [260–264] but there has been little experimental work on the trade-off of parasitic elements, particularly at very short gate lengths.

A great deal of work has been done on parasitic elements associated with bondpads and circuit elements further from the device intrinsic region. Particularly, the parasitic effects of gate feeds and airbridges [235, 265–268] have been considered, with some fairly involved processes for gate feed and pad airbridges resulting. The number of fingers and the layout used in devices has also been investigated in some depth [230, 261, 269, 270], but this is an area which begins to impede on microwave circuit design rather than intrinsic device issues.

6.4 HEMTs in digital applications

The scaling limitations of silicon, particularly the required oxide thickness, have led to a resurgence of interest in the use of alternative materials in recent years. III-V channels have clear transport advantages in comparison to silicon, with their extremely high electron mobility making III-V systems desirable n-type candidates. Until recently, however, the lack of a suitable native oxide for the unpinning of the Fermi level largely precluded the use of III-V FETs as MOS devices.

Since the discovery of techniques [271] for the reliable unpinning of the GaAs interface, various groups have investigated the transformation of the HEMT into a viable logic device. Of particular note are groups working towards the realisation of III-V MOSFETs based on HEMTs, where the cap is undoped and an oxide in place on the surface. Considerable research effort has been expended into the establishment of device operation, and transferring the known high performance of high-indium channels to a MOSFET context [272]. Early low-indium InGaAs devices have shown promise [14, 15], but realisation of a suitable oxide on high-indium devices remains immature [16, 17].

Additional to HEMT-based MOSFET work, some groups [11, 250, 273–275] have recently investigated the use of HEMTs as digital devices in their own right. Though competitive in terms of gate delay and drive current, particularly when using high-indium channels and short gate lengths, the key obstacles to HEMT logic remain, inescapably, related

to the use of a Schottky gate contact. The challenges for HEMTs as logic candidates therefore remain repeatable threshold voltage control, gate leakage and acceptable subthreshold slope [273]. Indeed, it is noteworthy that recent papers on this subject [11, 273] make use of etching or thermal sinking processes to reduce barrier thickness. Though these and similar [222, 276, 277] processes may yield enhancement-mode devices, their repeatability and hence logic suitability is questionable.

6.5 Performance simulation

The simulation of the potential performance of HEMTs has also received some considerable effort. Most methods [208, 278–286] have used physics-based Monte Carlo techniques to reveal carrier transport in the device, though some have used analytical methods [287–290]. Most work has been done to determine the likely methods of channel transport in these devices for process iteration, or modelling for circuit design.

Much Monte Carlo simulation work was done by Kalna, et al. at Glasgow [90, 103, 291, 292] to investigate HEMT scaling, with key findings underscoring the need to minimise the influence of parasitics and scaling of the complete device structure, not simply lateral scaling of the gate.

Some recent work has been undertaken by Ferry, et al. [22, 107, 108] to project the potential ultimate high-frequency performance of $In_{0.75}Ga_{0.25}As$-channel HEMTs if scaled using full-band Monte Carlo methods. Using a device structure strikingly similar to that proposed in later chapters of this thesis, it was shown that HEMTs might be expected to achieve up to 2.9 THz cutoff frequencies. It is noteworthy that these models reflect a partial picture of the devices, incorporating axial source and drain resistances, but excluding contact or gate resistances, any parasitic capacitance and leakage effects. As a result, this cannot be regarded as an achievable figure. The work, however, underscored the need for reduced source-drain separations in both reducing access resistances and enhancing the electric field, and particularly in scaling the barrier thickness proportionally to the gate length.

6.6 Summary

This chapter has reviewed the current state of HEMT-related research, covering fabrication-driven aspects such as approaches to gate lithography, efforts in improving the materials systems by a variety of approaches, optimisation of the device geometry and reduction

of parasitic contributions. Current approaches to device design have been discussed, in addition to the recent particular interest in HEMTs and HEMT-based MOSFETs for digital applications. To conclude, the potential performance estimates drawn from simulations of HEMT structures was discussed. The work described in Chapters 7-9 is based on the foundational work of this preceding research.

7. Development of sub-25 nm HEMT processes

7.1 Introduction

Chapter 3 has reviewed the principles underlying HEMT operation, and has highlighted the performance benefits associated with the reduction of the physical dimensions of the intrinsic region of a device, assuming the performance is not deteriorated by parasitic effects or incorrect scaling.

Chapter 4 has outlined the techniques available for the fabrication of nanoscale structures on semiconductors.

This chapter describes the application of these processes to the development of gate modules for the fabrication of HEMTs with critical dimensions as small as 10 nm. The processes developed had the aim of robust fabrication with the simultaneous reduction of parasitic device components.

7.2 Single-step gate processes at Glasgow

The standard method for T-Gate definition, as discussed in Section 4.6, is to use several layers of resist in a single exposure, such that a more sensitive top resist is more exposed than an underlying, less sensitive resist, giving a T-shaped profile in the resist bilayer.

Much work has been carried out at Glasgow in the past on the development of short-gate length T-gates. MESFET [293, 294] work in the late 1980-90s made use of the bilayer/trilayer PMMA/P(MMA-MAA) copolymer processes [295], whilst HEMTs as small as 80 nm were fabricated using a copolymer process [296].

In order to fabricate devices at shorter gate lengths, a multi-layer stack of resists comprising very high sensitivity ratios was used; principally a base layer of PMMA to define the gate foot and a layer of the high-sensitivity CAR, UVIII, to define the gate head [159, 297–299]. A layer of aluminium was initially used to separate the two resists to prevent intermixing, though this was later replaced with LOR resist [203], which allowed greater processing flexibility.

This resist stack was used extensively to routinely fabricate devices as small as 50 nm [300–302] using a Leica EBPG-5HR electron beam lithography tool with a minimum spot size of around 11 nm. Structures as small as 30-40 nm were further realised on planar bulk GaAs [203]. Smaller gates were more reliably fabricated using this resist stack when a dielectric mechanical support layer [205, 303] was added above the gate foot, where the gate foot was used for the dry etching of silicon nitride or silicon dioxide, followed by recess etching of the cap layer and metallisation.

A bi-lithography strategy [303, 304], discussed in further depth in the coming sections, was also investigated [204], and simulations suggested the possibility of fabricating smaller structures using this technique.

By a combination of the bi-lithography and dielectric support processes, gates as small as 25 nm have been fabricated. Functional HEMTs, however, were never fabricated.

7.3 Fundamental lithographic limitations

There are several fundamental issues in defining T-gates with critical geometries below 30 nm using a single lithography step and multiple thick resist layers of differing contrast. As discussed in Section 4.3.2, electron scattering effects dominate the definition of small patterns in resist by electron beam lithography.

Single-step T-gate processes use thick layers of resist, required to define the head of the gate, as described in Section 4.6. As a result of forward scattering effects, the spot size at the foot resist is considerably enlarged.

Consequently, whilst single-step processes will benefit from electron beam lithography systems with smaller spot sizes, scattering will, regardless, necessarily dominate the ultimate feature size produced.

To illustrate, the Monte Carlo simulation package, CASINO [305], was used to simulate

the scattering of 1000 incident electrons accelerated from 100 kV during the exposure of a resist bilayer of 50 nm and 300 nm PMMA. These resist thicknesses are similar to the PMMA/LOR/UVIII stack used for T-gate fabrication at Glasgow. As shown in Figure 7.1, even with an incident spot size of 4 nm, scattering effects produce electron exposure over a spot diameter of greater than 100 nm at the substrate, with the highest density within 30 nm of the centre.

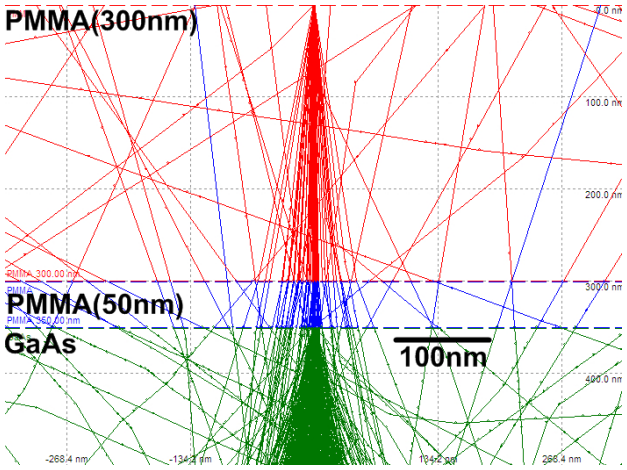


Figure 7.1: CASINO Monte Carlo simulation of the exposure of a single-pixel feature using a 4 nm spot at 100 kV in a composite bilayer T-gate resist.

As a consequence, it is to be expected that as lithography tools with smaller spot sizes are employed, forward scattering effects will account for an increasing fraction of the total exposure.

The Vistec VB6-UHR-EWF used in this project has a minimum spot size of approximately 4 nm, and is generally run at an accelerating voltage of 100 kV. As such, it represents one of the highest resolution commercial lithography tools available at present and an excellent platform with which to investigate the issues surrounding short gate length transistor fabrication.

A single-step T-gate exposure requires that the gate head and foot receive different doses, such that the dose to expose the gate head is insufficient to expose the foot, whilst the foot is exposed using a much higher dose by exposure through the gate head resist. The standard method for this exposure is generally the use of a "conditional figure assignment" (CFA) process, similar to that used to assign doses in proximity effect correction. The areas of the shape to be assigned different doses are assigned different

layers in the pattern file, though, alternatively, the dose assignment can also be achieved by feature size sorting. Each layer is then assigned a dose multiplier. The layers therefore receive a relative dose in ratio to their dose multipliers as a factor of the base exposure dose. The CFA step is incorporated into the pattern at fracture stage.

During exposure, the beam is stepped around the pattern area and the dwell times calculated using the relative doses defined by CFA. As a consequence, the complete shape with its multiple doses is exposed in a single step within a field. This method ensures that there is negligible drift of the stage and avoids the need to align one dose region to the other.

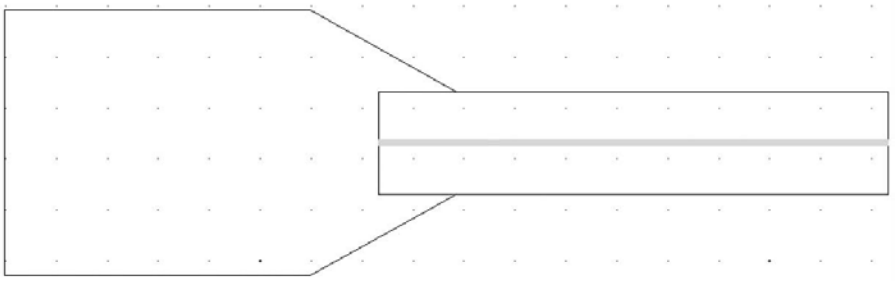


Figure 7.2: Designed layout of a single-step T-gate structure. The gate feed structure is also shown.

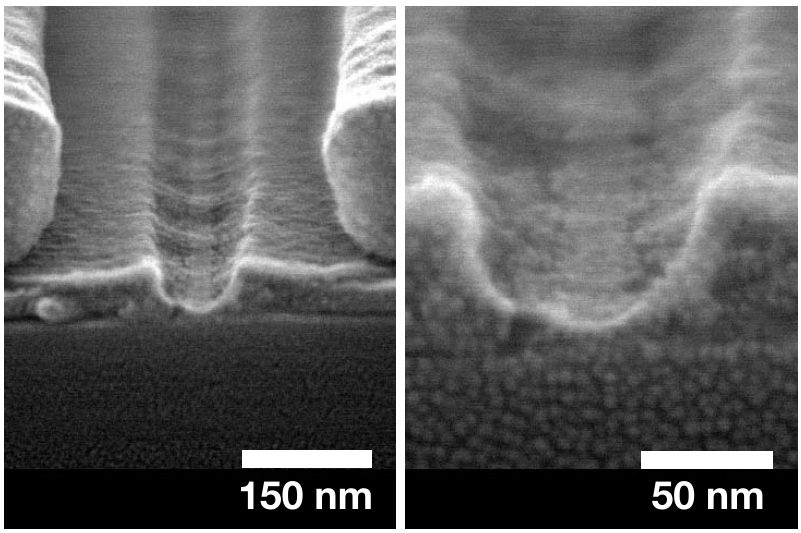
As a consequence, therefore, the T-gate is defined in the pattern file as a short area defined on one layer for the foot, and a larger area for the gate head. The general layout is shown in Figure 7.2. The foot length of the gate will then be determined by the dimensions of the central line (light grey), whilst the length of the gate head, which can be separately controlled, is determined by the dimensions of the larger shape. As a result, the foot length can be varied independently from the head length. The foot exposure will therefore be a factor of both the drawn dimensions of the central line and the dominant electron scattering processes.

The smallest gate lengths are achieved using a single-pixel line, where the beam is not stepped along the gate length, only along its width. The number of beam steps contained within a shape is specified by the beam step size, which is determined in software prior to exposure and, together with the exposure dose, determines the frequency at which the pattern generator is required to run. It is noteworthy that the beam step size can be defined independently from the spot size. As a result, spots may be set to overlap to a greater or lesser extent, or not at all. In the extreme case, where the beam step size is

much greater than the spot size, dots can be defined in the patterned area where a solid shape was desired. To ensure that a single-pixel line results, the drawn shape must be smaller than the beam step size, such that only one pixel results along the length.

To further explore the effects of scattering on the feature sizes achievable in T-gate resist with the VB6, the PMMA/LOR/UVIII resist stack which had proved reliable for the fabrication of 50 nm T-gates was spun on planar bulk GaAs and single-pixel and two-pixel lines were exposed using the 1 nA ($\sim$ 4 nm) spot at a 5 nm beam step size. The PMMA was 50 nm thick and the UVIII approximately 350 nm, with the LOR parting layer around 40 nm.

The two-pixel lines were then dose-tested between foot doses of 3700-5267 μCcm^{-2}, whilst the single-pixel lines received doses of 3700-8550 μCcm^{-2} .

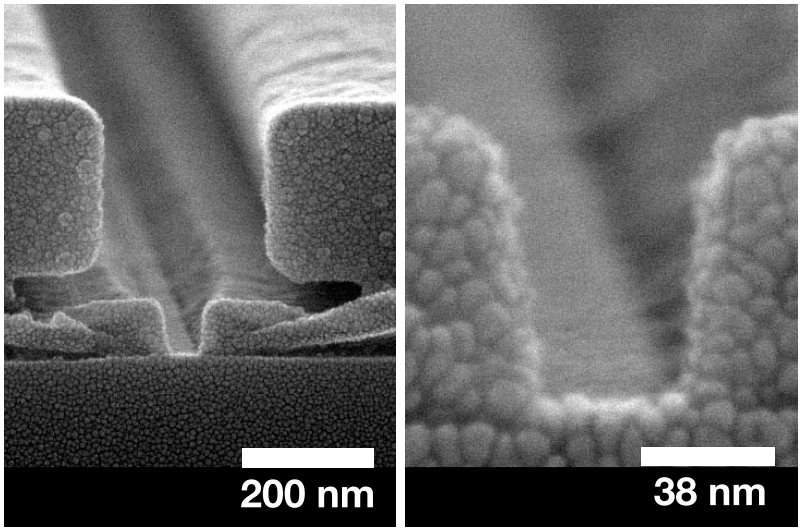


(a) Low magnification. (b) High magnification.

Figure 7.3: High-dose single-pixel exposures of PMMA/LOR/UVIII gate resist using a 4 nm spot at 100 kV. Though small features can be produced, contrast is poor as a result of scattering.

Extremely high dose values were required to develop out the gate foot using single-pixel exposures. At sufficient doses, features as small as 25-30 nm were produced, as shown in Figure 7.3, but as a consequence of forward scattering, the foot contrast was compromised, giving a large gradient to the sidewalls, unsuitable for metallisation. Indeed, the exposure profile of the resist appears to be a laterally-spread Gaussian distribution. Even at such high doses, some resist residue is also still evident.

In contrast, lines designed to be 10 nm, exposed using a 5 nm beam step size and hence requiring a two-pixel line, produced well-cleared features as small as 35 nm at much lower doses around 5000 μCcm^{-2} (Figure 7.4(b)). In addition, the use of larger designed features produced greatly improved foot resist contrast over the single-pixel exposure.



(a) Low magnification. (b) High magnification.

Figure 7.4: Two-pixel exposures of PMMA/LOR/UVIII gate resist using a 4 nm spot at 100 kV and a 5 nm beam step size. Contrast and uniformity is much improved over a single-pixel exposure.

These results can be interpreted as an indication of the scattering limitations inherent in a single-step gate lithography process. When using a small spot, the incident energy density is tightly constrained to the central line; hence the effect of forward scattering through the thick head resist is to increase the variance of the energy distribution to a greater extent than for larger exposure areas.

The probability of electron scattering is relatively constant for a given accelerating voltage and the recoil distance of a scattered electron shown in the Monte Carlo results of Figure 7.1 is relatively large. This is manifested in small lines as an increase in the energy distributed across the variance of the Gaussian exposure distribution, with a corresponding relative drop in energy at the exposure axis. Since the central exposure point requires an increased dose to develop completely, a corresponding increase in the exposure of the adjacent regions also occurs. The result is very poor contrast. Metallisation of such a profile would create mechanical instability and reduce lift-off yield.

In the case of multiple-pixel exposures such as that of Figure 7.4, the exposure Gaussians for the two pixels overlap within the central exposure region, creating a more even energy distribution across the foot exposure, with the energy distribution of a single Gaussian deposited in the adjacent regions to the patterned area. An increased energy density is hence deposited in the central region of the gate feature with respect to that of the adjacent, nominally-unexposed regions, resulting in superior contrast.

As a consequence, it is reasonable to conclude that single-step gate processes remain a lithographically viable solution for gate features on the order of 30 nm, where relatively large exposure areas are used, either by the use of multiple pixels or larger spot sizes. As exposure areas shrink, however, scattering effects become increasingly dominant, challenging the viability of these processes for high-yield fabrication.

7.3.1 Bi-lithography strategies

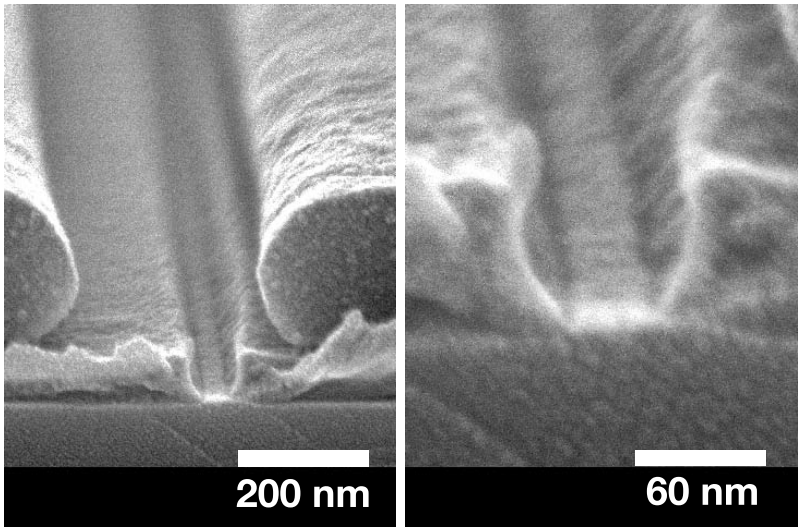
One strategy previously employed [204, 304] to counteract the scattering problems associated with single-step gate exposures is the use of two separate exposures to define the gate head and foot. By this process, the gate head area is exposed and developed, with no exposure of the gate foot at all. Given the relative resist sensitivities, the foot should remain unexposed at the head doses used. With the head resist removed in the exposure area, the sample is then re-loaded into the lithography tool, and the foot exposure carried out, requiring excellent lithographic alignment.

As a result of the two steps, electrons exposing the gate foot do not suffer from the extreme forward scattering imposed by the thick head resist; instead incurring only scattering from the thin gate foot. The foot can then be developed as usual.

This strategy was attempted using the 4 nm spot on the Vistec VB6, using the same development times as previously for doses of 2200-101000 μCcm^{-2}. At doses around 20000 μCcm^{-2}, gate lengths around 30 nm were produced, as shown in Figure 7.5.

The resulting contrast is superior to the standard single-step processes of Figure 7.3, though the foot profile remains far from vertical. There are likely to be several reasons for this.

Firstly, the doses involved are considerably higher than those required for the single-pixel standard process, since, as a result of the reduced scattering, exposure remains confined to a smaller area. As a result of the long dwell times, therefore, considerable energy will



(a) Low magnification. (b) High magnification.

Figure 7.5: Single-pixel bi-lithographic exposures of PMMA/LOR/UVIII gate resist using a 4 nm spot at 100 kV where the head and foot are developed separately. Contrast is somewhat improved over the standard process. The alignment is poor on this sample.

still be deposited in the adjacent gate regions, albeit with a much smaller range than for thick resist. Secondly, the effect of the head exposure, though relatively small, is not negligible in this process. The dose of around 100 μCcm^{-2} used to expose the head will to some extent expose the foot resist as well. This exposure will add to the exposure due to scattering from the foot dose, further eroding the contrast. Features smaller than 25-30 nm were not produced by this method; hence the primary benefit is improvement of contrast rather than the reduction of the achievable feature size.

One further point to note is the poor alignment of foot to head seen in Figure 7.5, which was seen in only some sample areas. It was expected that this may be related to sample tilt, with worst alignment results around the edges of the sample. It is therefore clear that an improved registration process is required for an alignment-based technique to be viable. These issues are discussed in greater depth in Section 7.4.4.

7.3.2 Other limitations of single-step processes

In addition to the lithographic constraints of single-step gate processes, there are multiple additional problems which affect the minimum feature sizes that can be effectively

realised.

Evaporation filling issues and mechanical stability

One of the primary issues with reduced foot lengths is the resultant increase in resist aspect ratio, which poses serious problems for effective metallisation. As the gate metal is evaporated, the metal builds up laterally over the resist feature, as for any evaporated structure. The problem is outlined in Figure 7.6.

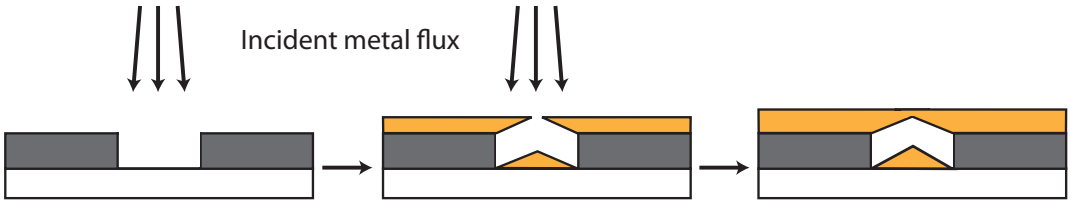


Figure 7.6: Schematic showing the process of pyramid gate formation during evaporation.

In the extreme case, which becomes more probable for increased aspect ratios, the metal deposited on the resist extends laterally by more than half the feature size before the desired metal thickness has been deposited in the exposed area. As a result, the exposed region actually receives a thinner metallisation than expected after liftoff, with a non-uniform profile across the length of the feature. As a result, lifted-off features in general have tapered edges. For increasingly short features, the obvious conclusion is that for high-aspect features, the metallised profile is likely to be dominated by the tapering regions, perhaps entirely. Short gates therefore frequently have a triangular cross-section, and are often termed "pyramid gates". The rate of lateral extension, unfortunately, is a factor of metallic grain size, mobility of the deposited metal in the evaporation environment and the incidence conditions. The metallised thickness and uniformity for high aspect ratio short features is hence extremely unreliable.

In the case of a T-gate resist stack, the problem is compounded since the metal is required to fill the foot resist to its complete thickness to allow connection to the gate head. Additionally, high aspect ratios in the foot are desirable to minimise parasitic gate capacitances from the large overhanging gate head, whilst etching the recess using the gate resist further increases the effective aspect ratio. If the combined aspect ratio is too high, the problem illustrated in Figure 7.6 can occur, disconnecting the gate foot from the head.

The problem remains severe even if the foot remains electrically connected to the head. The foot is required to remain robust enough to support the bulk of the gate head, whilst the top of the foot feature will be considerably less than its designed gate length as a consequence of the high-aspect filling issue. A gate foot which tapers too severely may hence be too short to support the bulk, and collapse during lift-off, regardless of its electrical continuity.

Various strategies [205, 306] have been developed to tackle this problem, mainly using dielectric support structures to support the gate head as the foot necessarily becomes increasingly tapered as its aspect ratio is increased. These processes generally make use of an RIE step to etch silicon dioxide or silicon nitride, adding complexity, whilst the exposure and evaporation processes may still be fragile or low-yield. The alternative to this approach is to sputter the gate metal, guaranteeing metal filling, since a sputter process is conformal, as outlined in Section 4.5.2. Sputtering, however, also potentially introduces damage to the underlying epitaxy.

Gate length extension

Using the gate resist to define the recess introduces a further problem in addition to exacerbation of the aspect ratio problem. After the recess step, an air gap exists under the evaporation mask, introducing shadowing effects as the metal flux deflects around the resist. As a consequence, the metallised feature physically extends under the resist mask, causing further profile tapering and, crucially, extending the gate length.

The effect is thought to contribute around 10 nm to the gate length of a 40-50 nm gate profile, requiring the realised gate resist profile to be smaller than the desired gate length [203]. The effect, however, will geometrically depend on the cap thickness; thinner caps will yield lower extension since the flux will deflect less into a smaller area. Regardless of the cap thickness, the effect is increasingly important at reduced lengths and may represent the minimum achievable gate length using this process.

7.4 Development of a two-step gate methodology

Surmounting these problems implies the development of a solution that circumvents the core issues. A solution is required which allows the optimum feature resolution to be achieved, without compromising the mechanical stability of the gate.

Achieving minimum feature size in resist requires the minimisation of scattering processes, primarily by thinning the resist; a strategy ideologically at odds with the single-step T-gate concept.

As a consequence, a two-step approach was developed to exploit the desirable resolution and contrast inherent to thin, single layer resist without incurring the mechanical instability and extension problems of single-step processes. Abandoning a single-step process additionally removes the most severe processing requirements of developer and resist compatibility, allowing optimal processes to be used.

As a result, a process was devised for the fabrication of extremely short gate features in a thin layer of resist, which is then transferred to a thin layer of silicon nitride. In a second lithographic step, a conventional T-gate structure is used to define the gate head and aligned to the gate foot, providing the bulk of the gate volume.

The process flow is outlined in Figure 7.7.

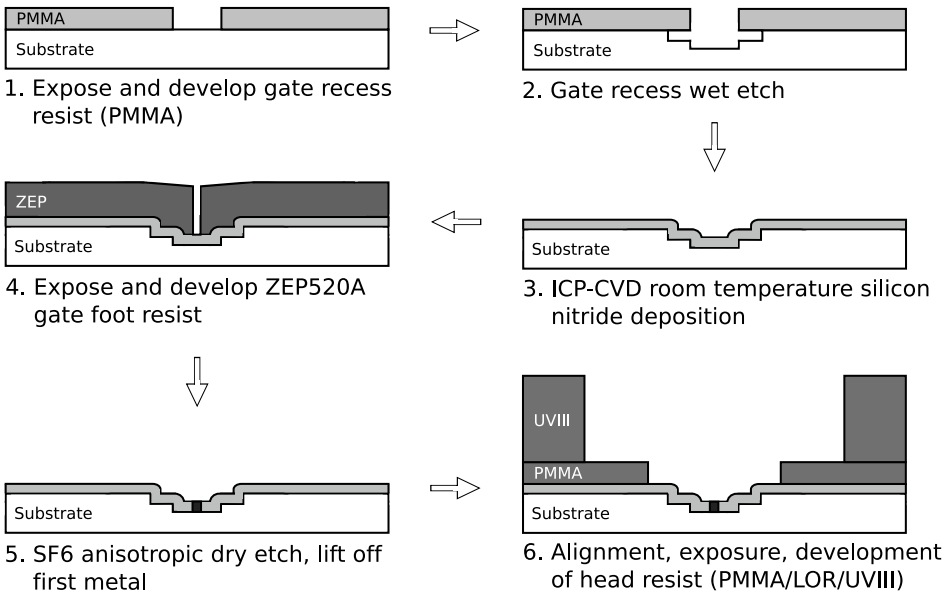


Figure 7.7: Process methodology for the development of sub-25 nm gates.

The move to this two-step strategy has several benefits:

- Firstly, it allows the use of optimally thin resist, minimising the lithographic limitations, and allowing the full potential of a next-generation tool to be exploited. The

pattern is then transferred to the silicon nitride using a low-damage RIE process.

- Using this process, the gate foot can be metallised separately from the gate head, circumventing the evaporation problems of high-aspect ratio features.
- Issues of mechanical stability are circumvented since the dielectric supports the gate bulk and the complete foot feature becomes metallised.
- The gate foot metallisation is constrained during evaporation by the silicon nitride sidewalls. There is therefore no air gap into which the evaporant metal flux can spread, minimising the feature size to that actually defined. In addition, the use of a dielectric pattern definition layer reduces the effects of resist flowing during evaporation.
- By performing the gate lithography after recessing, the complete recess trench is encapsulated in silicon nitride, conformally deposited at room temperature using ICP-CVD. The surface is thus protected from the ambient environment, a process often known as "passivation", frequently omitted from short gate-length research. Silicon nitride is ideal for this purpose, since it is an effective encapsulant of low permeability, can be deposited at room temperatures with minimal surface damage and easily removed in low-damage anisotropic dry etch processes.
- The upper lithographic level provides the bulk of the gate volume. The upper level is formed using a traditional single-step gate process to produce a double-tiered T-gate structure. This allows a large gate bulk to be realised, but with the gate bulk elevated as far from the surface as possible to reduce capacitance whilst maintaining mechanical stability. The upper level can then be lifted-off separately from the foot, resulting in complete metallisation of the double structure.
- All gate dimensions can be individually controlled.

7.4.1 Selection of gate resist

When considering the choice of resist for ultra-high resolution lithography and subsequent dry etching, four factors are of key significance; the ultimate resolution limits of the resist, the contrast achievable between developed and undeveloped areas (sidewall steepness), dry etch resistance and lack of damage to the underlying surface. Since the intention is to create as short an etched feature as possible, it is critical that the resist is capable of transferring that pattern optimally to the underlying area by maintaining its originally

developed features for the duration of the etch. As a result, the chosen resist must exhibit good resistance to the etch chemistry used, and the contrast must be as high as possible.

The most common high-resolution positive electron beam resist is PMMA, which has been reported to exhibit sub-10nm feature sizes [149, 307, 308] on silicon. As discussed in Section 4.3.2, however, it has poor dry etch resistance, and contrast can become poor at very small feature sizes.

Experiments were also carried out on PMMA using thin films, short development times, dilute developer, low-temperature development and ultrasonic agitation; all of which are claimed to improve achievable resolution [147, 307–310]. Despite these approaches, it proved impossible to reliably fabricate PMMA trenches of less than 25 nm on GaAs under any developer conditions at any dose using a single-pixel 4 nm spot. Therefore, also considering its poor etch resistance, PMMA was deemed inappropriate for the gate foot application.

The smallest feature sizes have been achieved using HSQ (hydrogen silesquioxane), with 5 nm features reported on silicon and diamond substrates [155, 311]. In addition, HSQ has excellent dry etch resistance; however, it is a negative resist, requiring a much greater area to be written to pattern the substrate, and, being similar in nature to silicon dioxide post-development, is impossible to remove in solvents. Since this process requires a trench to be created in silicon nitride and the resist then removed leaving only the gate metal and dielectric, HSQ is unsuitable for use in the proposed gate strategy, though it was successfully employed to create 30 nm T-gates [312].

ZEP520A is a high-resolution positive-tone electron beam resist which exhibits excellent contrast and superior dry etch resistance, whilst features as small as 10 nm [154] have previously been reported. Methods for its removal in solvent exist, and it seems the ideal candidate resist for this application.

A single-pixel dose test was performed with the minimum spot size using a 100 nm-thick layer of ZEP520A spun on a planar GaAs substrate and developed in o-xylene.

Even for such a relatively thick film, it was possible to define sub-25nm features, as in Figure 7.8. Extremely high contrast was also achieved; an ideal result for its use as a dry etch mask for the silicon nitride layer.

Consequently, ZEP520A was chosen to define the gate foot layer. Its considerable promise for use in defining smaller features is also noteworthy.

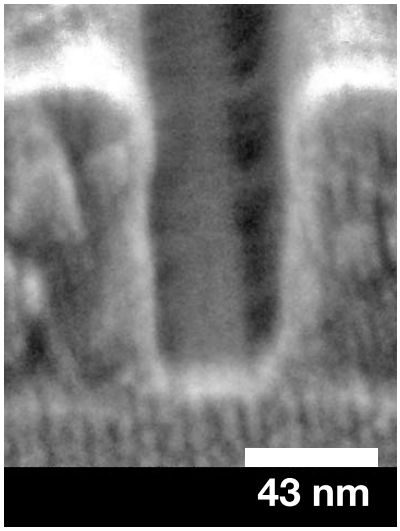


Figure 7.8: Resultant profile after single-pixel exposure of ZEP520A using the 4 nm spot and 30 s development in o-xylene. Note the resist profile has been slightly modified by SEM, causing upper resist shrinkage.

7.4.2 Silicon nitride processing

It was then necessary to find high-resolution processes capable of transferring this pattern into the silicon nitride. Any processing was required to be virtually damage-free, since it was necessary to process directly on the InAlAs barrier layer to which a Schottky contact would later be made, as a consequence of defining the footprint post-recessing.

ICP-CVD deposition of silicon nitride can be achieved at room temperature using zero sample bias; hence it provides a method of depositing a high-quality film without concerns over excessive surface damage, as evidenced by its use in previous sensitive III-V MIM capacitor and MOSFET work [256, 313, 314]. Additionally, the method is conformal, completely coating the etched and unetched surface. A nominal thickness of 50 nm was chosen as an initial starting point as a compromise between resultant capacitance and processing requirements.

As the silicon nitride layer is thinned, the capacitance between the 70 nm "foot" of the upper gate level and the substrate will increase. Considering the complete gate in a distributed fashion where the total parasitic gate capacitance is considered as a parallel sum of the various capacitive elements of its geometry (Figure 7.9), this will increase total gate capacitance. Conversely, as the nitride layer is increased in thickness, the

resist required to etch it must also be thicker; detrimental to the achievement of minimal feature sizes as discussed. As a result, there exists a trade-off between gate capacitance and gate length.

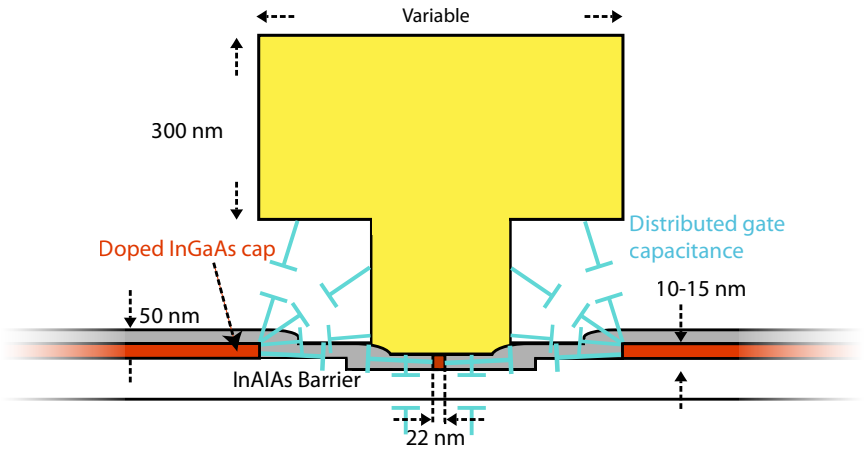


Figure 7.9: Schematic of capacitances arising from gate geometry.

The deposited dielectric layer is variable using this scheme. A silicon dioxide layer would yield a lower dielectric constant (3.9 for SiO_2 versus 7.5 for Si_3N_4 at room temperature) and hence reduced parasitic capacitances, but it is more difficult to deposit and etch in a damage-free manner. Silicon nitride is accordingly a more obviously suitable candidate dielectric for exploratory experiments.

Processes for the low-damage etching of silicon nitride were previously in existence for sidewall spacer etching for self-aligned III-V MOSFETs. SF_6 etches silicon nitride, but is relatively isotropic, creating non-vertical structures. The addition of nitrogen [256] introduces anisotropy, yielding vertical sidewalls. The process had been extensively studied [256] for the effects of pressure and particularly forward RF power and d.c. bias on the sheet resistance of GaAs MOSFET material, similar to the HEMT material used for this work, and shown to be effectively undamaging for RF powers of 20W or less.

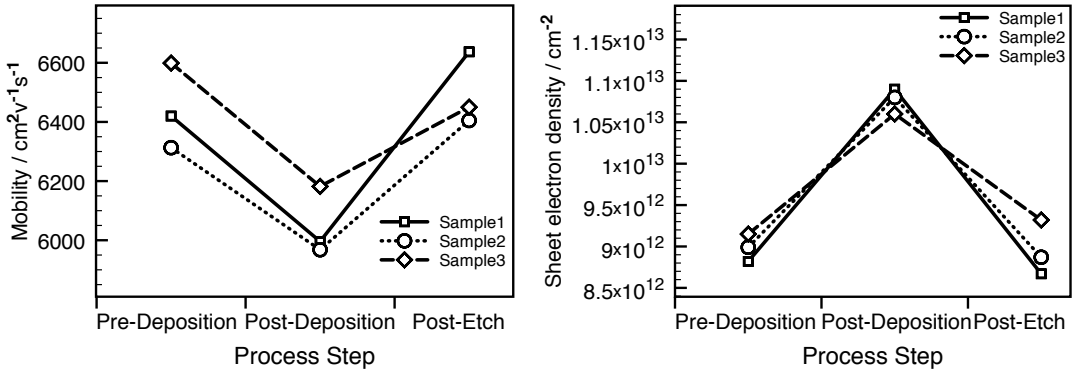
Plasma-induced damage

These processes, since already well-established, were selected for further analysis on HEMT material to be used for the definition of short gate-length devices.

It was crucial to establish their suitability for the purpose of deposition and etching of

silicon nitride on the sensitive InAlAs barrier surface without causing serious damage to the device material. Characterisation cells were fabricated on HEMT material, and 300 nm of silicon nitride was deposited at room temperature using the ICP process, far thicker than required for the gate process. A thinner deposition and etch would require far less exposure to the plasma so the thicker film deposition and etching represents the processing extreme.

The film was then removed using an unmasked "blanket" SF_6/N_2 process using reflectometry to determine the etch termination point.



(a) Mobility during plasma processing.

(b) Sheet electron density during plasma processing.

Figure 7.10: Transport measurements obtained from van der Pauw technique on capped HEMT material ($In_{0.8}Ga_{0.2}As$ channel) during processing of a 300 nm silicon nitride film.

The van der Pauw technique was used to measure mobility and sheet carrier concentration before and after deposition and etching of the 300 nm film. Three samples were processed in parallel to allow a general trend to be drawn from the measurements. The results are shown in Figure 7.10. It is interesting to note that mobility drops by around 10% following film deposition, whilst sheet electron density increases by around 15%.

Following etching of the film, however, both values return to around their pre-deposition values. The mechanism for the transport changes will be discussed in a later section, and is not fully known. Since the modifications occur only whilst the film is in place, returning to the original values following its removal, it is reasonable to conclude that the effect is a consequence of a modification of the semiconductor surface itself, rather than any damage-related mechanism.

Equally importantly, the values indicate that neither deposition nor etching processes intrinsically damage the underlying material: an indication of their suitability for use in device processing where the cap remains in place.

The proposed process flow, however, requires silicon nitride to be deposited both on the InAlAs barrier and the capped InGaAs as shown in Figure 7.9. As a result, it is equally important to determine any damage incurred on the recessed surface as a result of the deposition process.

In order to investigate these effects on the recessed surface, realistic thicknesses from 10-100 nm of silicon nitride were deposited onto recessed van der Pauw structures, where the cap layer has been selectively etched in succinic acid to expose the InAlAs barrier. Prior to silicon nitride deposition, the electrical properties of the material were measured. Changes in mobility, sheet electron concentration and sheet resistance were then calculated from their initial values. The transport measurements were then compiled against film thickness and any trends investigated. The results are shown in Figure 7.11.

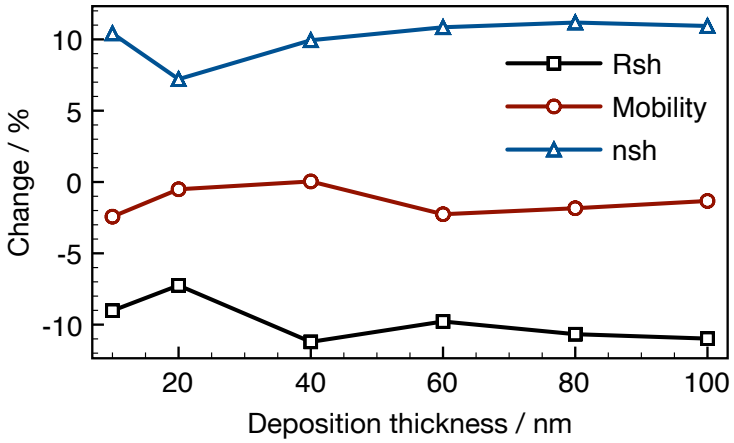


Figure 7.11: Changes in transport metrics following deposition of various silicon nitride film thicknesses on recessed van der Pauw structures.

Across the complete spread of deposited thicknesses, in contrast to the scenario on capped material, an enhancement in sheet electron concentration occurs, with a corresponding drop in sheet resistance, but there is no significant decrease in mobility. Though slight peaks are noticeable at 10-20 nm films, and the general trend is for slightly reduced enhancement with decreasing film thickness, the overall picture is a 5-10 % increase in electron density with effectively zero change in mobility.

The result is significant. These changes are consistent with a reduction of the surface potential in the gate recess region immediately adjacent to the gate, implying filling, passivation or other modification of the surface states, both in InGaAs and InAlAs. In the case of InAlAs, population enhancement has occurred with no net loss in transport efficiency. This in particular has implications for the device performance, where increased electron density implies improved drain current, with a potentially enhanced transconductance if the device's pinchoff characteristics remain unchanged.

In addition, any passivating effect on the surface states immediately adjacent to the gate is immensely significant, since, as discussed in Section 3.6.2, surface states are thought to play a major role in the kink effect, as well as effectively extending the gate length [106].

As a consequence, since the recessed region is very significant in determining HEMT transport, it appears that the effects of deposition will predominantly be to improve transport, effectively reducing channel resistance.

Silicon nitride etching

Since the SF_6/N_2 process [256] was confirmed to be undamaging, it was a suitable candidate for pattern transfer of the sub-25 nm gate foot into the deposited silicon nitride layer.

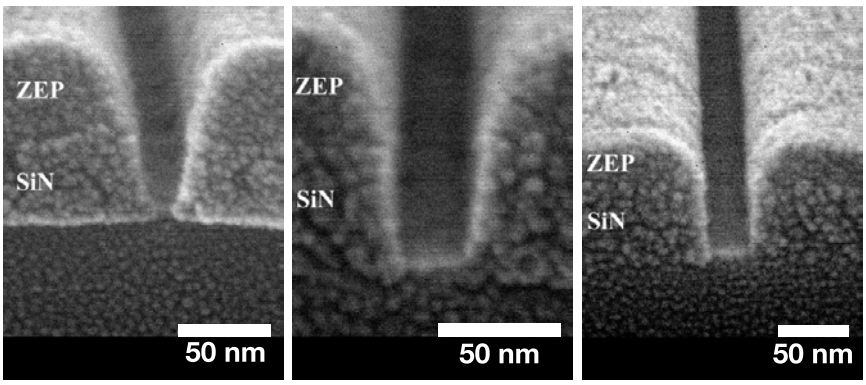
The ZEP520A resist film previously used was 100 nm thick. Taking into consideration the transport enhancement figures noted in Figure 7.11, it was decided to etch the gate foot into a 50 nm-thick silicon nitride layer. The reasoning was two-fold; 50 nm represents an optimal point where mobility is least affected but electron density is enhanced, and a 50 nm film can be etched using a 100 nm-thick resist mask.

50 nm of silicon nitride was therefore deposited on a bulk GaAs sample by ICP-CVD, and 100 nm of ZEP520A spun on top. Single-pixel lines were then exposed in the resist using the 4 nm spot as previously, and the dose varied across a grating of lines from 1700-7000 μCcm^{-2}. The sample was then exposed to an SF_6/N_2 etch chemistry for varying times.

Etching very small features requires longer than large areas, since the reactant gases have limited capacity for diffusion at the surface. As a consequence, "over-etching" is required past the apparent end-point of the etch process as determined by reflectometry.

Since the features produced in the resist for this application are extremely small, it was recognised that the over-etch time would be extremely significant, and the time required to completely etch the film much greater than the large-area time. As a consequence, it was decided to investigate various fixed etch times to determine the optimal over-etch time. An etch rate of around 16 nm/min is expected for this etch system.

The etched gratings were then cleaved and sputtered for cross-sectional SEM analysis. It is clear from Figure 7.12 that the etch time indeed plays a crucial role in the feature size achieved.



(a) 4m 15s. (b) 4m 45s. (c) 5m 15s.

Figure 7.12: Silicon nitride etch results in SF_6/N_2 RIE process for various etch times.

An etch time of 4m 45s gave optimal results; shorter times did not completely etch the nitride layer (Figure 7.12(a)). Longer times, though apparently producing more vertical sidewalls, eroded the resist edges and produced a slightly enlarged feature, as can be seen in Fig.7.12(c). As a result, the 4m 45s etch time was established as an optimal baseline for constructing a device process flow.

Each of the etched trenches in the dose grating was examined in cross-section by SEM and compared to the dimensions of the resist mask prior to etching, as shown in Figure 7.13. It is noteworthy that the resultant etch trench is uniformly around 2 nm smaller than the original resist for all exposure doses and in all cases the film is etched through completely. As a consequence, it can be reasonably concluded that both the pattern definition and etch process exhibit high uniformity and reproducibility: suitable for high-yield device fabrication.

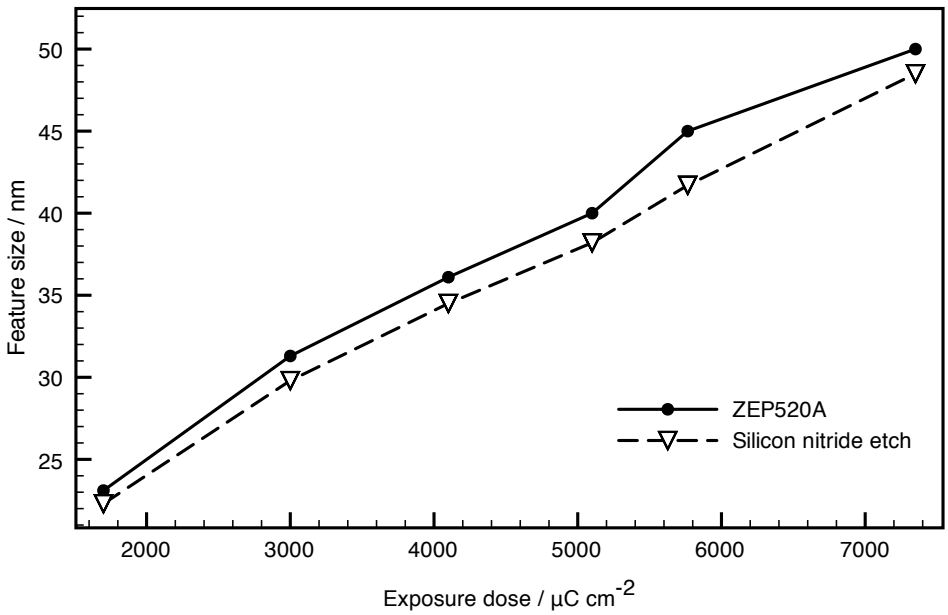


Figure 7.13: Variation of dimensions of resist mask and etched trench in silicon nitride with exposure dose following 4m 45s etch time.

Metallisation

It is clear from Figure 7.12 that for all SF_6/N_2 etch times, some residual ZEP520A remains; the resist is not completely eroded during the etch process. As a consequence, it is conceivable to use the remaining resist as a mask for the lift-off of an evaporated metal recipe to comprise the gate foot. Since one major problem with evaporating short gate length, high aspect metallisations is complete filling of the evaporant metal, metallising the gate foot separately circumvents a major fabrication issue.

Though conceptually ideal, some problems persisted; ZEP520A is infrequently used as a lift-off mask and in comparison to PMMA is much harder to strip in solvent [201], though it remains soluble, unlike a negative resist. Unfortunately, as with most resists, use as a plasma etch mask induces polymerisation, further reducing solubility. As a consequence, ZEP520A becomes increasingly difficult to remove after the silicon nitride etch completion.

Stripping in acetone has historically proved unreliable, requiring an overnight soak. As a consequence of its removal difficulties in acetone, previously ZEP has been removed using an oxygen plasma etch. Neither solution is ideal, since the resist must be removed

completely without inducing any damage to the sensitive barrier. In addition, a process-compatible solvent was required to enable ZEP520A as a useful resist for gate foot lift-off to enable the envisaged process flow.

Various solvent and agitation combinations were investigated on etched silicon nitride/resist samples for removal of the resist.

Firstly, the sample was soaked overnight in hot (50°C) acetone and inspected. A second sample was placed in an ultrasonic bath in hot acetone for 20 minutes. It was found that neither treatment completely removed the resist, leaving attached sections around the etched areas. Literature [315, 316] suggested the use of a commercial remover, Microposit 1165, which is an aggressive solvent blend based around n-methyl-pyrrolidone, but which is acetone-free [317]. The same samples were placed in hot Microposit 1165 remover for identical overnight and ultrasonic treatments as pursued for acetone removal, rinsed in de-ionised water, then inspected again.

The samples showed further resist removal, with areas of complete removal, but adjacent areas where partially-removed resist appeared to have re-adhered to the substrate.

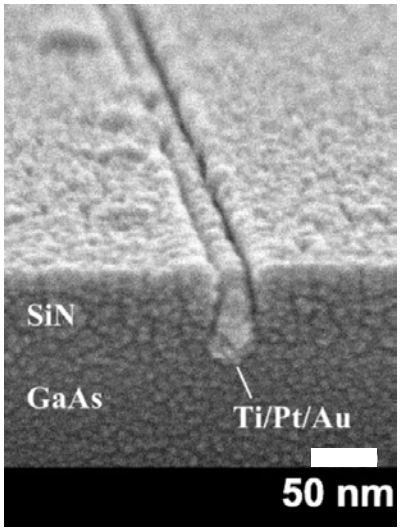


Figure 7.14: Profile resulting from lift-off of 15nm Ti / 15 nm Pt / 15 nm Au evaporated metallisation into 50 nm deep SF_6/N_2-etched silicon nitride trench. The resultant structure is virtually planar.

Two new samples were soaked in Microposit 1165, without the prior acetone treatment, then inspected. Samples subjected to both a short one-hour soak in hot Microposit 1165

and to a 20-minute ultrasonic soak showed complete removal of the ZEP520A. Shorter times later proved to also completely remove the resist. Since samples previously exposed to acetone were unaffected by exposure to Microposit 1165 remover, it was concluded that the acetone had an effect akin to cross-linking on sections of the resist, exacerbating problems with their removal. As a result, exposure to Microposit 1165 remover alone proved an effective method for stripping ZEP following RIE.

Since an effective solvent removal method had been found, the use of lift-off processes was enabled. A thin gate metallisation of 15 nm titanium / 15 nm platinum / 15 nm gold was evaporated into an etched 22 nm trench in 50 nm-thick silicon nitride and lifted off in Microposit 1165. The resultant profile is shown in Figure 7.14. As is clear from the figure, the substrate is left virtually planar with complete removal of the resist. In addition, problems metallising the high-aspect ratio feature appear to have been resolved and the trench is completely filled by the evaporated metal, though the characteristic taper discussed in Section 7.3.2 remains.

The process appears to have eliminated the metallisation issues of short-gate length fabrication: disconnection from the upper structure, feature spreading during evaporation and mechanical instability. The foot definition further provides a virtually planar surface for the upper gate lithography.

7.4.3 Upper gate lithography

As previously described, the upper gate level was envisioned as comprising a single-step T-gate process as previously fabricated in the department. The PMMA/LOR/UVIII system [203] provides great flexibility in lithography for this type of structure, and so was chosen to realise the upper level.

The purpose of the upper gate was to contain the bulk of the gate cross-sectional area, whilst presenting a minimal footprint on top of the etched silicon nitride as shown in Figure 7.14 and elevating the gate bulk as high as is mechanically feasible. The combination of a short footprint and high gate head should result in reduced parasitic capacitances, shown in Figure 7.9.

A compromise was selected where the footprint could be minimised, whilst remaining a feasible lithographic prospect in relatively thick PMMA, realising the elevated gate head. A further consideration is that the upper level must be aligned to the lower etched level with relative margin for misalignment. In order to achieve these requirements, a 70 nm

single-step process was selected. This was defined as a double-pixel exposure of a 10 nm designed "foot" using a 5 nm beam step size. The upper "head" was initially defined as 300 nm long.

A trilayer of 100 nm 8% 2010 PMMA / 50 nm (1:4) LOR / 300 nm 58% UVIII was spun on planar bulk GaAs and exposed at a range of doses up to 6000 μCcm^{-2}. The sample was then plasma ashed for 60 s at 40 W.

The sample was then cleaved and sputtered for cross-sectional investigation by SEM. The doses were then characterised and the resultant foot dimensions measured. A foot dose of 4200 μCcm^{-2} gave a 70 nm footprint in 100 nm-thick resist, meeting the upper gate requirements specified.

7.4.4 Alignment

The gate strategy chosen relies on the excellent alignment capability of modern electron beam lithography tools. As a result, determining the actual capabilities of the system was crucial to the successful realisation of structures by the processes envisaged.

Alignment accuracy is determined by the methodology used to measure the position of the first lithographic level and the errors in that process. As discussed in Section 4.6.1, metallised or etched markers are generally used to provide known points from a prior lithographic steps to which subsequent levels can be aligned. The usual method for locating these markers, and the method employed in the VB6 is a "mark locate" strategy which aims to find the edges of the previously defined marker.

The electron beam is scanned over the expected marker location, previously defined in software and the backscattered electron detector is used to form a profile of the backscattered electrons across the marker [142]. In the case of metallised markers formed from dense metals, electrons are increasingly backscattered relative to the background substrate level. As a consequence, the profile formed shows the marker position as a region of increased backscatter intensity. In the case of etched markers, spikes in backscatter intensity occur at the etched edges. These profiles are illustrated in Figure 7.15.

In conjunction with the stage positions at the edges of these differential intensity regions, the marker positions can be determined, but are strongly influenced by errors in the specified or actual marker dimensions or position errors. Such inaccuracy can arise as a consequence of substrate tilt or rotation, and hence propagate into the next level of

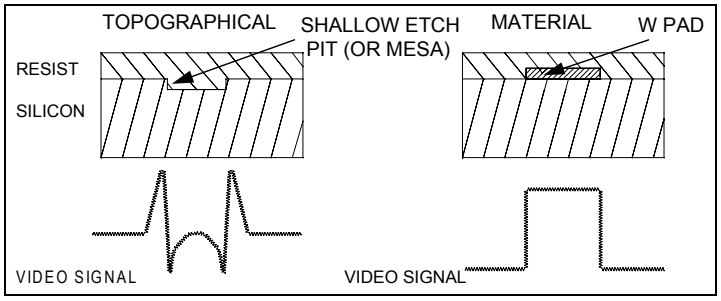


Figure 7.15: Backscatter profiles used in EBL mark locate routines. Adapted from VB6 operator manual [142].

lithography as misalignments to the first level. The inaccuracy is exacerbated over large marker separations, and areas outwith the marker-bounded areas exhibit deteriorated alignment. The use of one marker in each corner is usual, though it is possible to use fewer.

To minimise these effects "cell" alignment strategies can be used, where an initial "global" alignment is performed as previously over a large substrate area, with a subsequent alignment performed over a smaller area, with correspondingly smaller marker separations and smaller markers. The positions of the markers are consequently subject to reduced errors from sample orientation and stage position, allowing their positions to be determined more precisely.

Consequently, cell alignment strategies generally yield improved alignment, but require additional time.

To measure the practical alignment capability of the system, a 15 mm $\times$ 15 mm sample was written with both global and cell markers defined in 20 nm titanium / 130 nm gold. Nine cells were spaced evenly across the sample, with each cell comprising a 3 mm $\times$ 3 mm area with a block of markers in each corner. Global markers were 20 μm $\times$ 20 μm as for the standard HEMT process flow (Section 4.6), whilst cell markers were 4 μm $\times$ 4 μm. On the same level, the first gratings of several small verniers were also written. The vernier technique relies on a sequence of deliberate misalignment, whereby the two aligned exposures determine the deviation from the ideal, hence the alignment accuracy. Verniers with grating offsets of 5, 10 and 20 nm were defined, with several sets of verniers placed in the global alignment area at even spacings across the sample, allowing any effects of substrate bowing, sample flatness and rotation to be determined.

Several sets of verniers were also placed inside each cell, spaced evenly across the cell area.

The second level of the verniers was then defined in a 4% 2010 / 2.5% 2041 PMMA bilayer and examined both optically and by SEM, as shown in Figures and 7.16 and 7.17.

Looking firstly at the global alignment verniers, alignment appeared to be better in the x-direction than in the y-direction. This was expected, since the beam had previously been determined to have an instability in the y-orientation which was not present in its x-axis; an instability attributed to seismic or electromagnetic interference. The global alignment accuracy measured using the verniers was around 40-50 nm in x, and as poor as 100-120 nm in y.

The cell alignment accuracy proved to be similarly better in the x-axis. Alignment in the x-direction was shown to be 15-20 nm as measured optically and by SEM, whilst y-alignment was somewhat poorer, around 30-35 nm. These measurements were repeatable and uniform across the sample area.

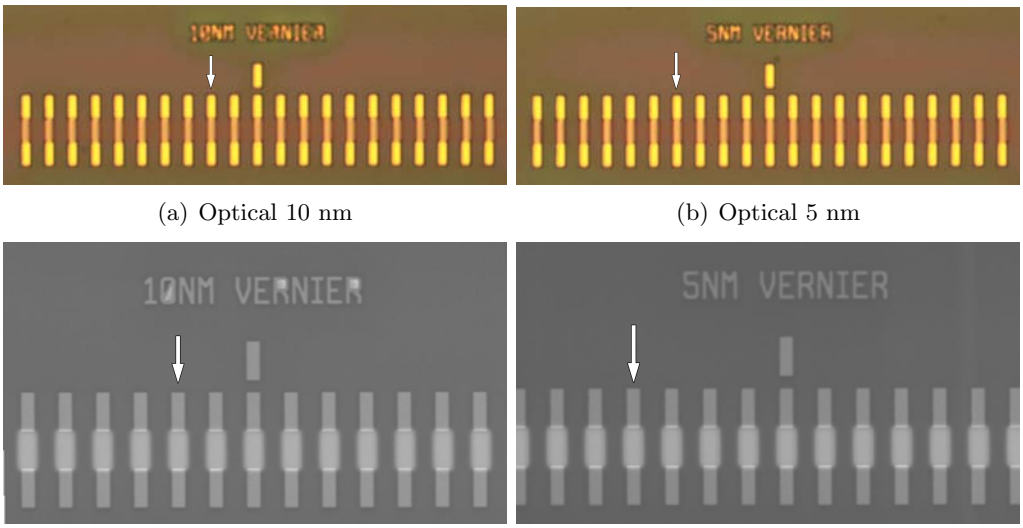
Given the requirement of the gate process for the alignment of features on the order of 20 nm to the centre of features on the order of 60-70 nm, a 35-40 nm misalignment was unsatisfactory, since shorting of the gate foot to the cap or disconnection of the gate head could result. Consequently, global alignment alone was insufficient. Cell-aligned y-alignment also remained insufficient. The most viable alignment strategy for the new gate process was therefore to capitalise on the x-alignment capability possible via cell alignment strategies.

As a consequence of this analysis of the alignment capabilities of the lithography tool, the decision was taken to use cell alignment techniques, with critical geometries (gate and recess) aligned along the x-axis.

7.4.5 Complete gates and resistance measurements

Complete 22 nm gates were fabricated on a planar GaAs substrate to determine the effectiveness of the gate process in producing robust short gates. Since the prime motivation for T-gate fabrication is resistance reduction whilst maintaining a short gate length, structures capable of probing for measurement were defined.

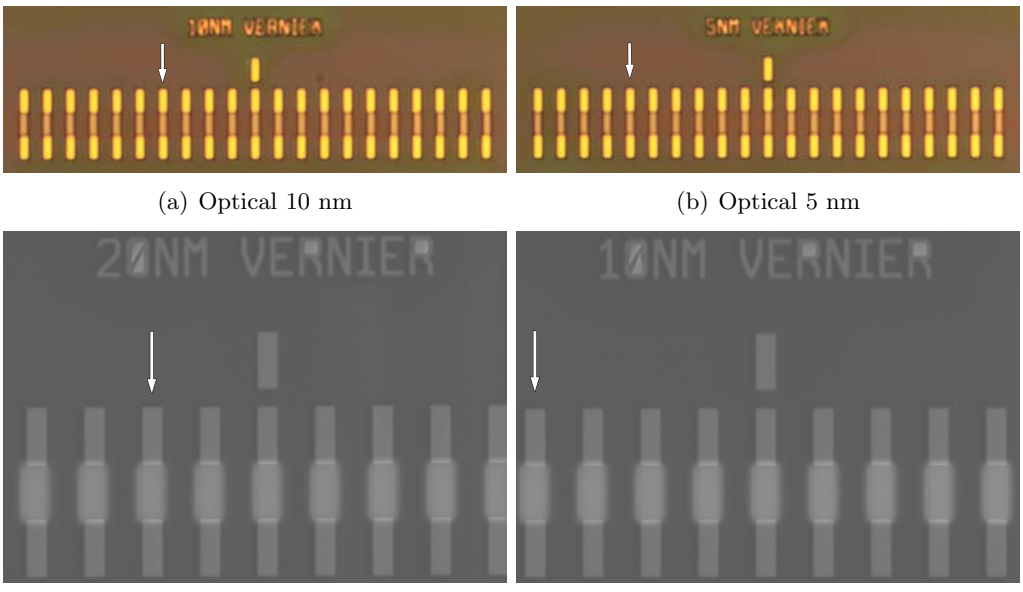
Markers identical to those described in the alignment test setup were defined in a first lithographic step. 50 nm of silicon nitride was then deposited and a 100 nm layer of



(a) Optical 10 nm (b) Optical 5 nm

(c) SEM 10 nm (d) SEM 5 nm

Figure 7.16: X-alignment verniers for cell-aligned region. Each period is an alignment step of the indicated offset, hence the aligned bars marked by arrows indicate alignment of 15-20 nm.



(a) Optical 10 nm (b) Optical 5 nm

(c) SEM 20 nm (d) SEM 10 nm

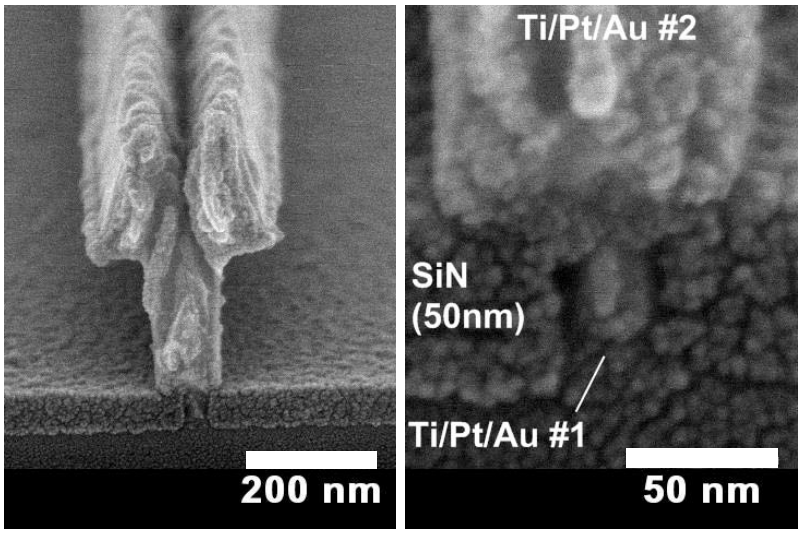
Figure 7.17: Y-alignment verniers for cell-aligned region, indicating alignment of 30-35 nm.

ZEP520A was spun, exposed and the silicon nitride etched for 4 m 45 s as described in Section 7.4.2 and lifted off with 15 nm Ti / 15 nm Pt / 15 nm Au. The gate width (the many-pixel large dimension) aligned parallel to the y-axis to capitalise on the x-alignment capabilities of the VB6.

The upper lithography process outlined in Section 7.4.3 was then aligned to the gate foot using cell alignment and exposed using the dose required to yield a 70 nm geometry. Various gate head dimensions were defined in UVIII to explore the structural limits of the system and allow the range of resistivities possible to be explored.

To complete the structures, thick (20 nm Ti / 200 nm Au) bondpad structures were defined in a 12% 2010 / 4% 2041 PMMA bilayer and lifted off. It is important to note that any processing occurring after upper gate lithography was carried out using resist bake temperatures of 120°C to prevent melting of the unsupported gold gate head.

Several of the structures were cleaved and sputtered for cross-sectional SEM, shown in Figure 7.18.



(a) Overview of completed gate. (b) Closeup of gate foot.

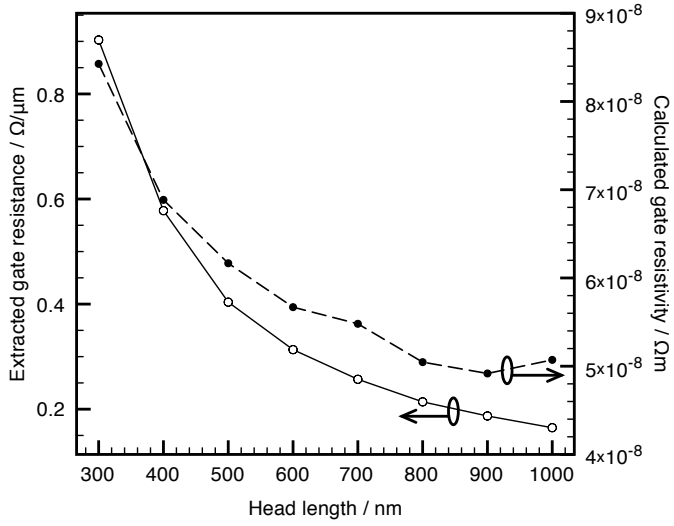
Figure 7.18: Cross-sectional SEM images of completed 22 nm gate, showing complete metal filling and 22 nm gate length. The gate foot is completely encapsulated in silicon nitride.

As shown in Figure 7.18, the gate alignment, though not perfect, was easily sufficient for the gate process. There is, additionally, complete metallisation of the whole gate feature.

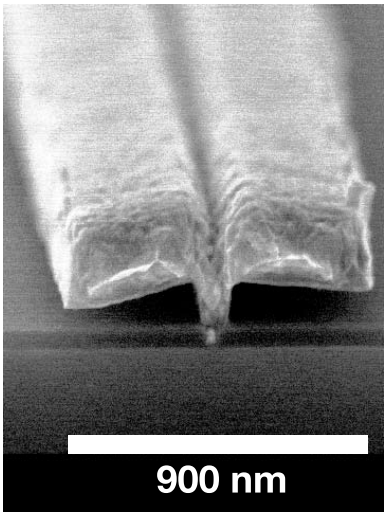
This was uniform across the test samples, and a mechanical gate yield of over 95% was achieved, confirming the robust nature of the process flow.

As the head and foot are defined in separate lithographic steps, there is considerable latitude in controlling the total gate resistance, with a corresponding trade-off in parasitic gate capacitance. To investigate the effectiveness of this system in reducing the resistance of a 22 nm T-gate, the gate foot was fixed at 22 nm, the middle footprint at 70 nm, whilst the upper head length was varied between 300 nm and 1 μm, (Figure 7.19).

Transmission line method (TLM). structures of the 22 nm footprint T-gates with widths over a range of 25-100 μm were fabricated for various gate head dimensions, including the usual gate feed structures described in Section 4.6.3. The Ti/Au probe pads allowed the resistance of the gate structures to be determined by a linear fit to the four-probe current/voltage traces. The measurements were then averaged across three cell locations to take account of any substrate tilt errors or resist thickness fluctuations which may affect the gate lithography. By extrapolating the measurements for each length, a figure for gate resistance per unit width can be extracted. The various gate resistances for each head dimension are shown as the solid line in Figure 7.19(a).



(a) Gate resistance and resistivity calculations for various gate head lengths from TLM measurements.



(b) SEM image of a 22 nm T-gate with a 900 nm head.

Figure 7.19: Gate resistance and resistivity calculations show the expected reduction in resistance with area and a drop in resistivity approaching that of bulk gold for increasing gate head dimensions.

By measuring the dimensions of the various gate structures by cross-sectional SEM, approximate areas for each of the gate geometries were determined. Resistivity values were calculated using the measured mean gate resistances for each, graphed as the broken line in Figure 7.19(a). It can be seen that the gate resistance decreases with increasing cross-sectional gate area, as expected. In addition, as the gate head length is increased, the resistivity of the gate also decreases, approaching the value for bulk gold of $2.2 \times 10^{-8}\Omega$m. This can be attributed to the decreasing relative compositional fraction of titanium and platinum, both much less conductive than gold, for gates of larger cross-sectional area, since the 22 nm foot section does not scale and hence contributes a constant resistance to each structure. This variable resistivity results in a non-linear relationship between the gate resistance and its calculated area.

It is therefore clear that the gate resistances of the 22 nm gates can be reduced to levels comparable to much longer gates; indeed the gate resistances are effectively dominated by the resistance of the gate head. The largest gate head dimension therefore yields resistivity akin to that of a 1 μm pyramid gate, whilst retaining a 22 nm footprint. These results verify both the mechanical stability of the gate module and its electrical effectiveness as a gate strategy.

Since the gate process had proved to be intrinsically damage-free, mechanically stable, high yield and low-resistance, it was determined to be suitable for incorporation into a device process flow.

7.5 Integrated device process flow

Though short gates had been successfully fabricated, their incorporation into a device process flow requires significant further development. As a result of the complexity of the new gate module, significant changes to the standard process flow were required.

It was anticipated to make use of the previously-developed non-anealed ohmic contact processes developed in the department [318], allowing the gate lithography to be performed on a locally planar area, previously shown to increase uniformity [177]. As a result, this was determined to be a sensible route for the fabrication of gates using the new process.

The other major changes to the process flow were due to the separation of the recess etch from the gate lithography, and the splitting of the gate lithography into two stages.

There are several fundamental requirements when fabricating devices that must be met. As with the generic HEMT process modules described in Section 4.6, the ohmic and isolation levels may be interchanged. The gate level incorporating the gate feeds and traversing the mesa must, however, be defined after the isolation level. In addition, any processing which occurs after a defined T-gate must be undertaken at low resist bake temperatures to prevent flowing of the gold, which has a low melting point, and to avoid any potential diffusion of the gate metal into the barrier, a phenomenon known as gate sinking which can result in a loss of control over threshold voltage.

It is critical to note that when nanometric features are defined by lithography, slight changes in the resist thickness, as a consequence of topographic variation or proximity to other structures, cause significant differences in the resultant exposure. As a result, dose testing performed on a planar substrate is largely irrelevant when applied to complete devices.

To circumvent these problems, it was decided to perform the first level of gate lithography immediately after the gate recess definition, minimising the variations in surface topology during the critical gate lithography which would otherwise arise from the presence of a device mesa or contacts in the immediately adjacent area.

7.5.1 Resist uniformity and sample topography

The use of ZEP520A also highlighted an additional issue during alignment testing. Thin films of PMMA and other resists commonly used flow very well over sample topography when spin-coated. In the case of ZEP520A, however, it proved impossible to maintain a uniform film thickness over pre-existing features such as the device mesa or metallic markers. Where the topographic gradient was shallow, the film thinned over the features; in the extreme case where the gradient was large, the resist did not flow over the features at all. A sample was spun with ZEP520A then examined by Atomic Force Microscopy (AFM) to analyse the resultant topography. As is evident from Figure 7.20(a), the resist, nominally 100 nm thick, does not flow correctly over a nominally 150 nm thick cross.

In the case of the alignment test samples described in Section 7.4.5, this became a problem during further processing. The process flow developed relies on a dual-processing strategy using a single exposure of resist, masking both the silicon nitride RIE and metal lift-off. In regions around the gold markers, the resist did not flow sufficiently over the topography, exposing the metallic surface and immediately adjacent substrate as the resist was eroded

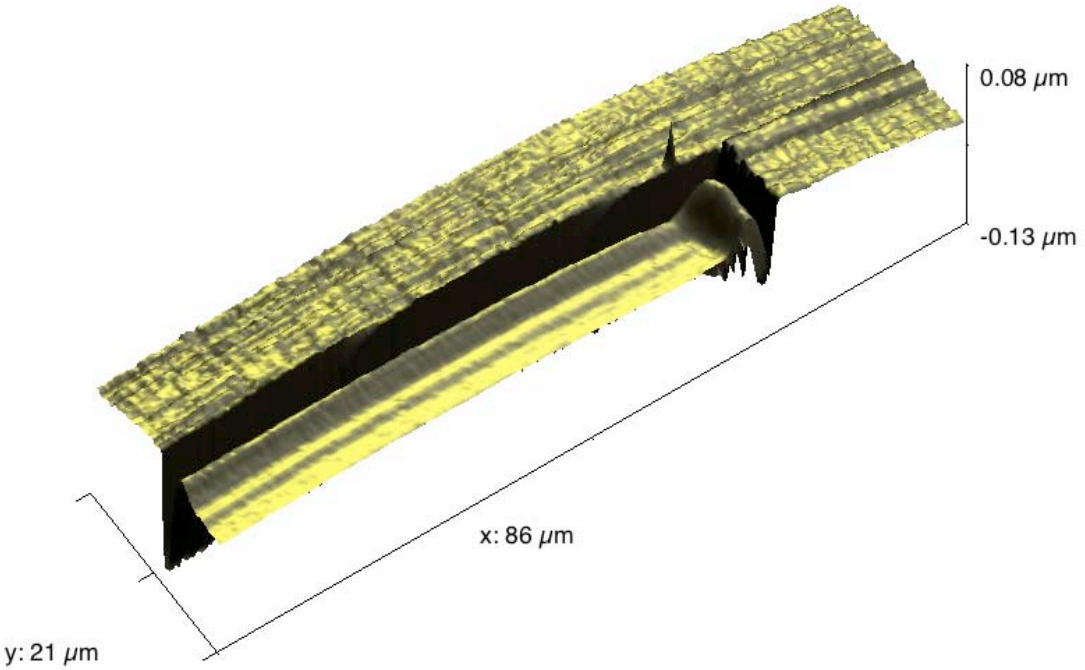
during RIE. As a consequence, the markers and substrate were exposed to both the RIE process gases and the subsequent gate metal evaporation, coating them in gate metal and rendering them ineffective for the precise alignment required, obvious in the optical micrographs of Figures 7.20(b) and 7.20(c). The issues with resist flow were assumed to be due to viscosity.

As a consequence of the topographic issues, an additional level of lithography was introduced into the process flow to protect the markers from damage. A single layer of thin PMMA was used to mask the markers before the ZEP520A was applied to the substrate. The alignment cells in which the devices would be placed were left exposed by this masking step, whilst the cell areas were several square millimetres in area. It was hoped this would have negligible effect on the ZEP520A film thickness, hence exposure doses.

In order to provide maximum flexibility for the definition of highly-scaled devices, it was also decided to capitalise on the two-step nature of the gate process to split the complete gate definition into is constituent parts, with intermediate process steps. The gate foot was defined on a virtually planar substrate to maximise yield and ensure uniformity. The second gate definition would limit the use of further high-temperature resist bake times, with corresponding reductions in achievable resolution due to the presence of additional solvent in the resist. It was anticipated that freedom in defining the ohmic contacts might be beneficial to optimise the short gate length devices, and so to allow lithographic freedom, ohmic definition can be placed between the gate steps or after both, dependent on the processing required. Since the gate was required to traverse the isolation mesa, it was also necessary to define the mesa before the complete gate was fabricated. The gate foot itself, however, had no requirement to cross the mesa if electrical connection was made by the head lithography. Since the upper gate level was shown to have uniform electrical contact to the gate foot (Figure 7.18), the mesa was defined to undercut the gate foot geometry only slightly, with the upper gate level, defined immediately after isolation, traversing the mesa.

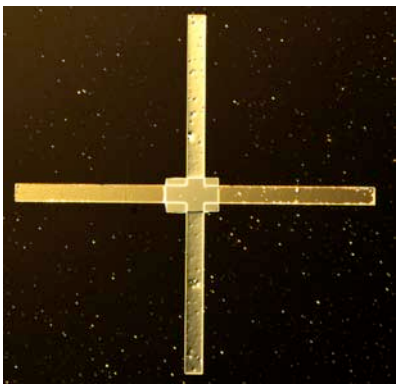
The optimal process flow for sub-25 nm devices therefore appeared to be:

1. Markers
2. Recess
3. Silicon nitride deposition
4. Marker protection



(a) AFM scan of resist flow around metallised cross. The resist thickness drops to virtually zero approaching the cross.

(b) Optical micrograph of gate metal adhered to marker level structures.



(c) Dark field optical micrograph of adhesion of gate metal to cross.

Figure 7.20: Lift-off problems with ZEP520A as a result of topographic variations.

5. Gate 1
6. Ohmic
7. Isolation
8. Gate 2
9. Bondpad

As a consequence of the early deposition of silicon nitride, the ohmic, isolation and bondpad levels were also required to include a silicon nitride RIE step to selectively expose the surface in the relevant regions as required whilst maintaining complete encapsulation of the active region of the device.

The key processing steps are outlined in Table 7.1 together with the resultant structures at each step. Silicon nitride is in place in the background after the Gate 1 step, then selectively removed in subsequent steps. The film remains in place between the gate and ohmic contacts, in the device intrinsic regions.

7.6 Initial material design

With the capability to fabricate short gate lengths verified, the design of suitable material was crucial.

Previous successful devices fabricated in the department have used a variety of material systems: lattice-matched and pseudomorphic devices on GaAs and InP, as well as metamorphic devices with channels grown both lattice-matched to the buffer or pseudomorphically on the metamorphic buffer. In particular, devices with the highest performance were, unsurprisingly, fabricated pseudomorphically on InP. The fastest devices reported at Glasgow had a cutoff frequency of 550 GHz [202], and comprised a 70% InGaAs channel grown pseudomorphically. Although the precise epitaxial dimensions were never published, the fastest metamorphic devices fabricated at Glasgow, which had a 440 GHz cutoff frequency, featured a 20 nm barrier. The fastest non-annealed devices, with an f_t of 490 GHz, had a 15 nm barrier. All sets of devices featured a 50 nm gate length.

As discussed in Section 3.8.4, vertical scaling of the device architecture with gate length is very important; hence, it is crucial that the aspect ratio of the intrinsic device is

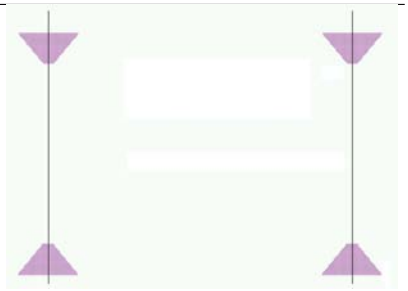
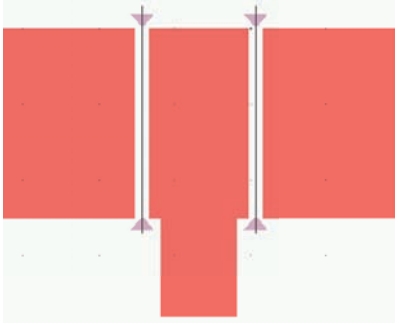
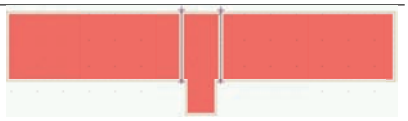
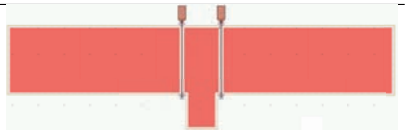
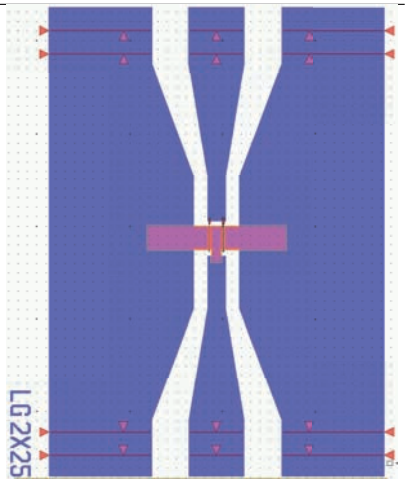
Step	Processing details	GDS overview
Recess	De-oxidise, succinic acid wet etch, silicon nitride deposition	
Gate 1	Silicon nitride etch, gate 1 lift-off.	
Ohmic	Silicon nitride etch, ohmic lift-off. Silicon nitride is etched under the contacts	
Isolation	Silicon nitride etch, orthophosphoric acid etch. Silicon nitride and semiconductor are removed outwith the mesa area.	
Gate 2	Gate 2 lift-off.	
Bondpad	Silicon nitride etch, bondpad lift-off. Silicon nitride is removed from under the bondpad area.	

Table 7.1: Critical device processing summary.

not detrimentally reduced during the scaling of the gate. The 50 nm devices previously fabricated featured aspect ratios (gate length : gate-channel distance) of 2.5-3.5.

It was therefore concluded that the design of initial material for the first devices fabricated using the new two-step process would be scaled based on these layer designs, but scaled only moderately in the first instance, with more aggressively-scaled designs to be realised after the successful demonstration of initial devices.

7.6.1 Non-annealed layer design

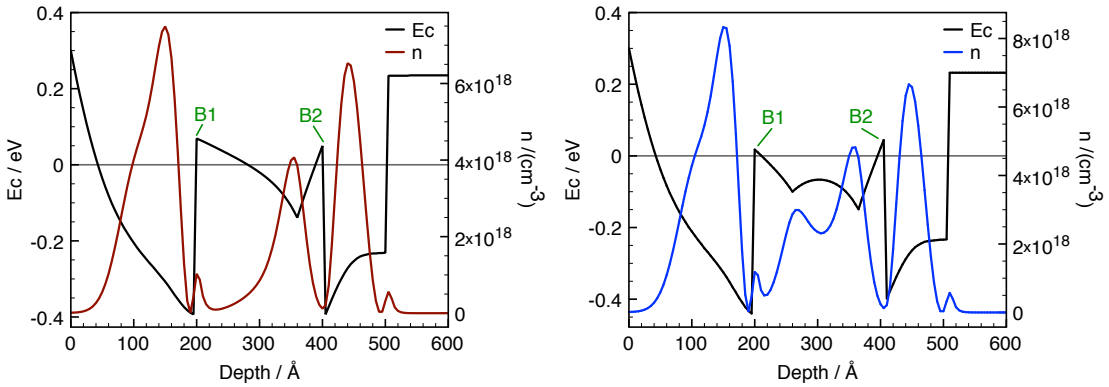
Previous work [319] has shown the importance of layer design in optimising vertical and lateral carrier transport through the ohmic contact regions, allowing low-resistance contacts to be formed without the need for an annealing step and therefore the ability to reduce the thermal budget of the fabrication process. This also allows the definition of the gate layer before ohmic deposition, improving uniformity and performance [177], which may prove to be of importance for high-frequency devices.

Designing material for use in a non-annealed system is complex, since such material must be engineered to minimise conduction band energy barriers to electron transport between the cap and channel. As a consequence of the use of multiple heterojunctions, the conduction band profile varies greatly with respect to the Fermi level. As a result, for an electron injected into the cap layer, barriers exist on its path to the channel in areas where the Fermi level is considerably below the conduction band. This is an important consideration with respect to the formation of parasitic access resistances. By careful placement of dopant and layer thicknesses, the variation of the conduction band with respect to the Fermi level can be controlled. Whilst a highly-doped cap is used to form ohmic contacts, without optimisation, significant energy barriers exist between the cap, barrier, spacer and channel. Previously, this band engineering has been achieved by the precise placement of multiple dopant planes, modulating the magnitude and width of these barriers [184, 318].

Simulation can be used to analyse the effects of variation on the conduction band profile. As described in Sections 3.4 and 3.4.1, conduction band profile and occupancy are specified by the Poisson and Schrödinger equations, which must be simultaneously and self-consistently solved. Snider's programme, "1DPoisson" [46] is one software solver which does this, given the appropriate material properties.

By defining simulation constants such as the surface potential and applied bias, the con-

duction band profile can be analysed for numerous layer designs, incorporating variations in layer thickness, composition, dopant density or placement. Graphing the resultant output yields a meaningful understanding of the conduction band geometry and resultant electron distribution, as shown in Figure 7.21.



(a) Unoptimised single delta structure.

(b) Optimised with second dopant plane.

Figure 7.21: Unoptimised and optimised conduction band profiles for non-annealed layer structures lattice-matched to InP. B1 and B2, annotated, are the key energy barriers to electron flow between the cap and channel.

Although the main purpose of the simulation is to determine a layer structure appropriate to low-resistance contact formation, it is also crucial to ensure suitable conductivity in the channel after the cap is removed. As a consequence, simulations are carried out both with the cap in place and without it, assuming a corresponding shift in surface potential from 0.3 eV for $In_{0.53}Ga_{0.47}As$ [58] to 0.65 eV for $In_{0.52}Al_{0.48}As$ [54] and no additional material changes. Unfortunately, the situation implies a conflict between optimisation of the vertical structure to minimise energy barriers under the cap and optimisation of channel transport. Some primary reasons are the additional ionised impurity scattering induced by the presence of additional dopant planes and the formation of parasitic subchannels in the barrier or buffer layers.

It is also useful to note that in the interfacial regions between the capped and recessed regions, assuming imperfect etch processes, the conduction band profiles are expected to vary from one simulation result to the other as a consequence of the partial etching of the cap. As a result, Fermi level pinning and hence electron dynamics in these regions are impossible to predict accurately.

Figure 7.21 shows the conduction band geometries and electron populations for a single-

doped HEMT structure lattice-matched to InP, featuring a 20 nm cap, 16 nm barrier and 10 nm channel. Figure 7.21(a) shows the unoptimised layer structure, where large conduction band barriers, labelled 'B1' and 'B2', protrude above the Fermi level between the cap and channel. Figure 7.21(b) shows an identical structure, but with an additional delta doping plane inserted 6 nm into the barrier. It is clear, without changing other doping levels or thicknesses, that the barriers 'B1' and 'B2' are greatly reduced.

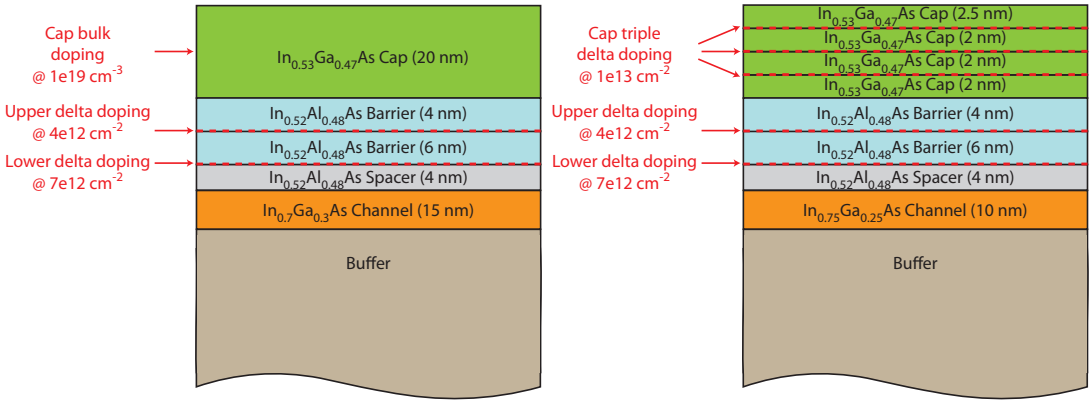
It is, however, worth noting that additional doping is not the only way to modify the band geometry. Variations in layer thickness, cap dopant and compositional fraction have significant combinatorial effects which, together, significantly affect the resultant bands.

It was decided to base the starting material on a previous pseudomorphic layer structure grown on InP, which featured a 15 nm 70 % indium channel and a 15 nm gate-channel separation, optimised for 50 nm gate lengths and yielding high performance devices. To increase the material performance, however, it was decided to improve the electron velocity characteristics in the channel by increasing the indium concentration of the InGaAs channel ternary to 75%. Although this will increase saturation velocity in the channel, it will also induce additional tensile strain in the layer, so the layer was also thinned to 10nm, as opposed to the 15nm used for 50nm pHEMTs.

Figure 7.22 compares the layer structures.

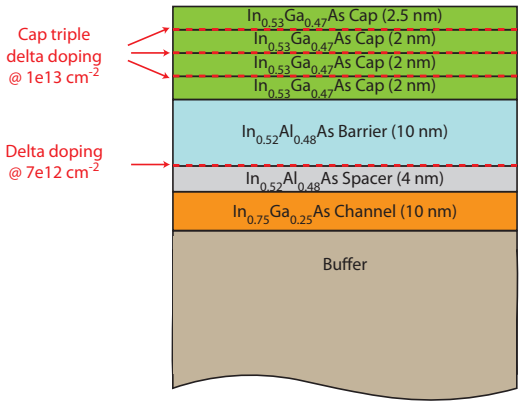
Previous double delta doped layers have featured one delta doped plane just above the spacer layer for modulation doping and one midway through the barrier layer, which reduces the magnitude of the barrier which forms between the cap and barrier layers, labelled 'B1' in Figure 7.21(a). Extending the doping strategy used for these 50nm devices to the new 75% material, however, creates a problem, since the depth of B1 becomes excessive due to the decreased bandgap of the channel (Figure 7.21(a)), which will increase non-annealed contact resistances.

For these new devices, therefore, a new solution was proposed. This method involved increasing the doping of the cap layer by the use of multiple delta doping planes inserted with a uniform distribution throughout the cap, circumventing issues of bulk dopant activation, which generally tend to restrict achievable doping densities. This increased cap doping has the effect of pulling the conduction band below the Fermi level, which, despite the large conduction band offset between the cap and barrier layers, can be enough to remove B1 entirely. The cap layer thickness was additionally scaled from 20 nm to 10 nm to ease lithographic concerns.



(a) A1941 - Original $In_{0.7}Ga_{0.3}As$-channel 50 nm material

(b) C216 - New double-delta doped structure (triple-doped cap)

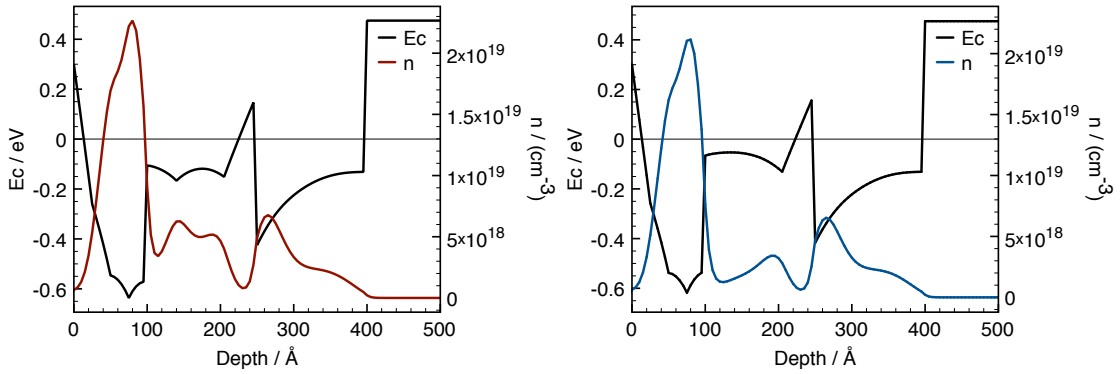


(c) C217 - New single-delta doped structure (triple-doped cap)

Figure 7.22: Layer structures of 50 nm material basis and new single- and double-doped layer structures.

As the cap doping concentration is increased, by increasing the density of delta doping planes throughout the cap, the barrier magnitude is progressively reduced until it drops below the Fermi level. This effect has a knock-on impact on the structure of non-annealed devices, since there is no longer such a strong requirement for the upper plane of delta doping to form ohmic contacts.

Given the degree of the attenuation of the upper B1 barrier, it seemed possible to remove the upper delta doping plane entirely, as per Figure 7.23(b), without compromising the access resistance characteristics of the device. Since the double delta-doped strategy has, in the past, produced excellent devices, and since the resultant high carrier concentration



(a) Double-delta doped structure

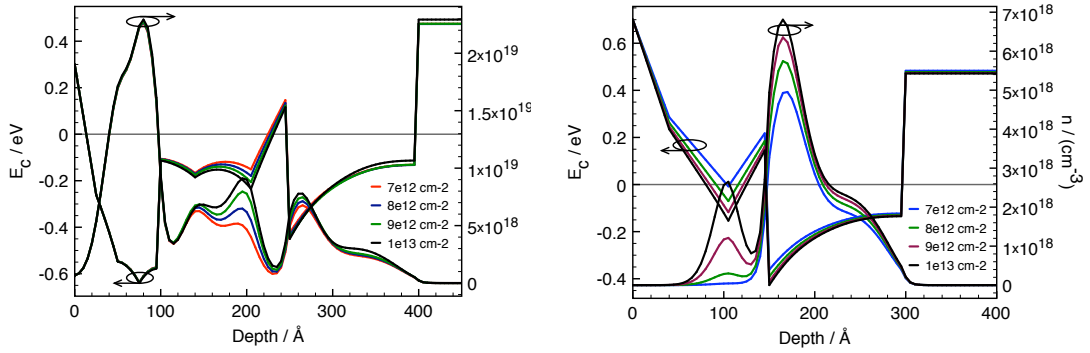
(b) Single-delta doped structure

Figure 7.23: Effect of cap delta doping on conduction band energy barriers.

can result in increased output currents and improved transconductance [291], it seemed wise to produce devices based on both the single and double doped strategies. Reducing the upper delta doping, however, should result in decreased scattering, impact ionisation and gate leakage effects [291]. In addition, reducing the overall doping should increase mobility due to decreased ionised impurity scattering. A comparison of a single-doped device with the established double-doped standard is therefore of interest.

By similar logic, it would also be possible to increase the lower delta doping concentration, in order to narrow B2 as seen in Figure 7.21(a). In order to investigate the impact of this effect, various simulations were carried out for doping concentrations up to $1x10^{13}cm^{-2}$, both with the cap on, required to minimise the access resistances, and with the cap removed for gate deposition. This second simulation step is important to ensure the correct distribution of electrons under the gate, which is essential for effective channel modulation.

Considering Figure 7.24 for increasing delta doping concentration, it becomes apparent that while for higher doping concentrations, the barrier is moderately thinned and decreased in magnitude, (Figure 7.24(a)), the electron concentrations under the gate in the recessed case redistribute such that electrons accumulate in the region of the delta doping at the point where the conduction band passes below the Fermi level, (Figure 7.24(b)). Since these electrons form a parasitic parallel channel in the spacer or barrier layer, they are subject to increased scattering, degrading performance: clearly an undesirable effect which must be eliminated.



(a) Capped (b) Recessed

Figure 7.24: Effect of varying lower delta doping concentration.

At concentrations where the electron distribution remains agreeable, the barrier is not significantly thinned for the ohmic contacts. As a result, it does not seem sensible to increase the lower delta doping concentration.

A further point of interest is to study the effect of gate recess depth. As part of a HEMT recess process, the cap layer is etched off in the gate region, which itself may result in some etching of the underlying barrier, or a short second recess etch can be used to form a two-step etched trench in which the gate sits. These processes etch the barrier layer, effectively altering the thickness of the barrier layer under the active region of the gate. By simulating the scenario with the cap removed, as previously with a thinned barrier, this effect can be evaluated. An interesting method is to perform the simulation for a device in which the lower delta doping concentration has been increased to the point where a parasitic channel has formed. The simulation was run with various etch depths from 0 to 5 nm and the effect on the carrier distribution noted, as can be seen in Figure 7.25.

The unetched material features a large triangular well at the delta doping which is considerably below the Fermi level. As a result, there is a considerable concentration of electrons situated around this dip. As the recess depth is progressively increased, however, the distance from the surface, where Schottky effects dominate the profile, to the delta doping decreases. Accordingly, the delta doping has a reduced effect on the conduction band, decreasing the depth and width of the well formed.

As a result, fewer electrons will accumulate in this region. As etch depths are increased, and indeed the etch progresses through the upper delta doping plane, the effect becomes

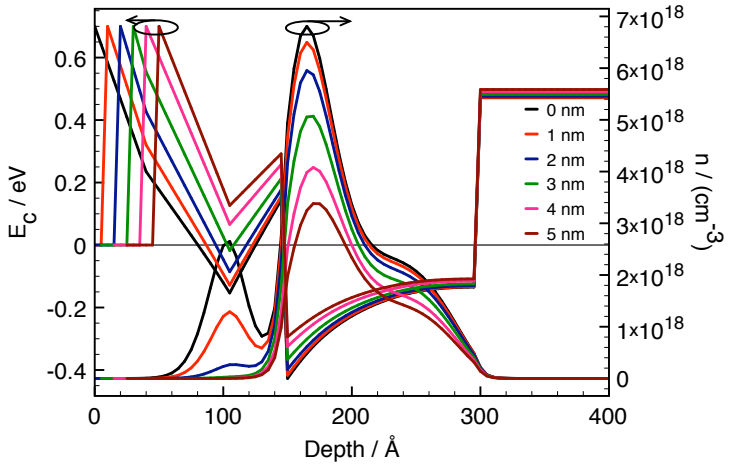


Figure 7.25: Band diagrams and electron concentrations resulting from varying etch depths at the second stage of a double recess process. As the etch depth is increased, the effect of the delta doping plane on the conduction band pinning becomes degraded. Plots are offset such that the channels coincide for all etch depths.

more pronounced, until for large etch depths, no well is formed at all.

Consequently, if the recess etch can be accurately controlled, it would be possible to use increased delta doping concentrations to reduce access resistances if required, then offset the parasitic channel formation problem locally around the gate region by the use of well-controlled recess etching. This could therefore be a useful technique in the design of the recess geometry of a device, ensuring that in designing the recess to minimise fringing electric fields around the gate region, the carrier distributions are not adversely affected.

Such precision could theoretically be achieved using "digital" etch processes [238, 239], which rely on the diffusion-limited oxidation steps and the subsequent etching of these oxides intrinsic to III-V materials (Section 4.5.1).

Given these results, both single and double-doped wafers C216 and C217 were grown by the department's MBE group as initial device wafers. Unfortunately, the department's InP growth systems were unavailable for use to grow the wafers; as a consequence, both were grown metamorphically on GaAs with the same layer structure.

7.6.2 C216 and C217 material characterisation

Both wafers were characterised electrically and to assess surface morphology: issues particularly of significance since metamorphically-grown devices in general have inferior surface properties which could cause difficulties for small-scale pattern transfer.

Transport properties were measured by van der Pauw techniques both with the complete structure and with the cap removed over the measurement area using the succinic acid selective etch process. The resultant sheet electron concentrations, resistivity and mobility are listed in Table 7.2.

	C216		C217	
	Capped	Recessed	Capped	Recessed
Sheet electron concentration / (cm^{-2})	6.27×10^{12}	2.16×10^{12}	4.56×10^{12}	1.72×10^{12}
Mobility / (($cm^{-2}v^{-1}s^{-1}$)	7223	11288	8724	11313
Sheet ρ / (Ω/sq)	138.1	256.5	157.1	321.8

Table 7.2: C216 and C217 transport properties measured by the van der Pauw technique.

As anticipated, C216, the double-delta-doped material, exhibited larger sheet electron concentrations than C217, both in capped and recessed structures. C216 in particular showed a channel electron population 26% greater than C217. Increased doping, however, should increase scattering probabilities, reducing mobility. Comparing recessed mobilities, however, that of C216 was reduced by less than 0.25% over C217. As a consequence, C216 exhibited a sheet resistivity 20% lower than C217.

Contact resistances to both structures were measured using the Transmission Line Method. Ohmic contacts were formed using an established Au/Ni/Ge recipe (11 nm Au/11 nm Ge/11 nm Au/11 nm Ge/20 nm Au/12 nm Ni/80 nm Au) [184]. In addition, the cap was removed between the ohmic contacts for one set of structures in proportion to the TLM gap, as in the structures of Figure 5.4. In these structures, a nominally identical capped region was in place for all structures, with the varying lengths of recessed region the transmission line variable under test.

From the TLM results, it is clear that in addition to superior transport properties, C216 also exhibited improved contact resistances. It is useful to note that the TLM results for sheet resistivity compare well with the van der Pauw measurements of Table 7.2. As such, it would appear that the extra layer of delta doping, in addition to providing extra carriers, significantly reduces the access resistances to the channel. The measurements

	C216		C217	
	Capped	Recessed	Capped	Recessed
Contact resistance / (Ω.mm)	0.12	0.18	0.15	0.46
Sheet ρ / (Ω/sq)	119.5	278.7	168.4	349.6

Table 7.3: C216 and C217 transport properties measured using the Transmission Line Method both in capped and recessed cases. Results are averaged over several uniformly-distributed sample sites.

for capped contact resistance are comparable, as would be expected from structures with nominally identical cap layers. The reduction in the case of C216 may be a result of additional unintentional doping from the delta-doping layers.

In order to properly assess the access resistances in the recessed case, a method for extracting real resistances from the combined current paths as portrayed in Figure 3.17 was required. From previous work carried out on non-annealed contacts, a suitable method had already been developed [184], using various recess lengths in a TLM layout, described in Section 5.3.

Using the previously-extracted capped resistivity figures and measurements obtained by AFM or SEM of the structures, the true access resistances can be extrapolated using the recessed TLM. For C216, the normalised extracted contact resistance was 0.18 Ω.mm, whilst C217 exhibited an extracted figure of 0.46 Ω.mm. The recessed case therefore portrays a much larger differential in contact resistance between the layer structures, where the contact resistances of C217 are over 2.5 times those of C216.

Given the superior contact resistances obtained from C216 and the marginal decrease in low-field transport over the single-doped case, it seemed that the double-doped strategy still yielded optimal performance overall. As a result, though in principle, single doping should be sufficient for contact formation, double doping was the preferred strategy.

Devices were fabricated on both C216 and C217 for comparison, though enhanced performance was expected from C216.

7.7 First-generation 22 nm device results and discussion

Initial devices were fabricated on both C216 and C217 using the 22 nm gate process flow, using the standard thick ohmic recipe usually used for HEMTs at Glasgow. 20 μm global and 4 μm cell markers were placed first to form 3×3 mm cells, with devices

aligned along the x-axis.

The recess level was then exposed in single-layer 2.5% 2041 PMMA and etched using a double-recess process involving selective succinic acid / non-selective orthophosporic acid [258]. First, a short succinic acid selective etch removes the cap. Secondly, a very short dilute orthophosphoric acid, non-selective etch removes the barrier in the area of the first recess. A longer succinic acid etch then extends the first recess trench laterally. Double-recessed trenches such as that in Figure 7.26 were formed, with the cap removed in a region around 150 nm long. In order to minimise the probability of misalignment of the gate to the recess, a longer final recess time of 60s was used to extend the first recess to 150 nm.

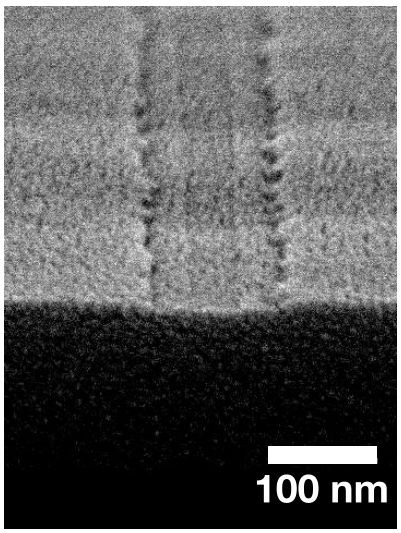


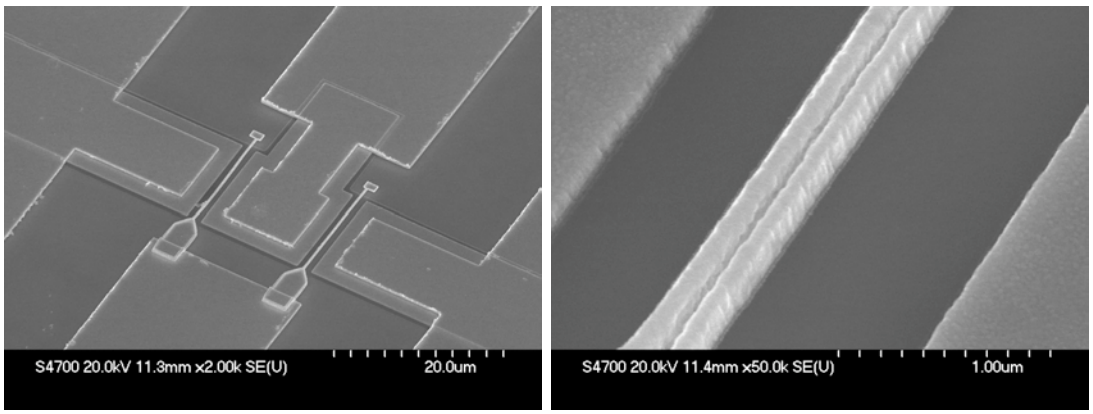
Figure 7.26: Double recess formed using 12s succinic/5s orthophosphoric/30s succinic acid processes. The lower recess is approximately 3-4 nm in depth, and is around 40 nm long. The upper recess is around 100 nm and removes the cap selectively.

50 nm-thick silicon nitride was then deposited by ICP-CVD. Marker protection resist was exposed in the device regions, but remained over the markers, then the ZEP520A gate resist was exposed at a dose of 1700 μCcm^{-2}, aligned to the etched recess trench. The gate was then etched for 4m 45s in SF_6/N_2 as previously. Prior to gate metallisation, a short hydrofluoric acid dip was used to deoxidise the InAlAs surface, as was standard practice in the department [320]. The upper gate was then aligned using an upper head length of 300 nm.

Following gate definition, ohmic contacts were defined with a standard spacing of 2 μm, the resist baked at 120°C to prevent gate degradation. The silicon nitride was etched under the contacts for 4m 45s to ensure complete removal. The ohmic contacts were then metallised using the standard ohmic recipe used for the previous TLM structures during wafer characterisation and were not annealed. Following ohmic liftoff, thick 50 nm NiCr / 400 nm Au bondpads were formed.

The complete process flow used is described in Appendix A.

Only 22 nm devices were fabricated in this batch, and the device widths were varied from 25-200 μm. Top-down SEM images of the completed devices were taken, and are shown in Figure 7.27.



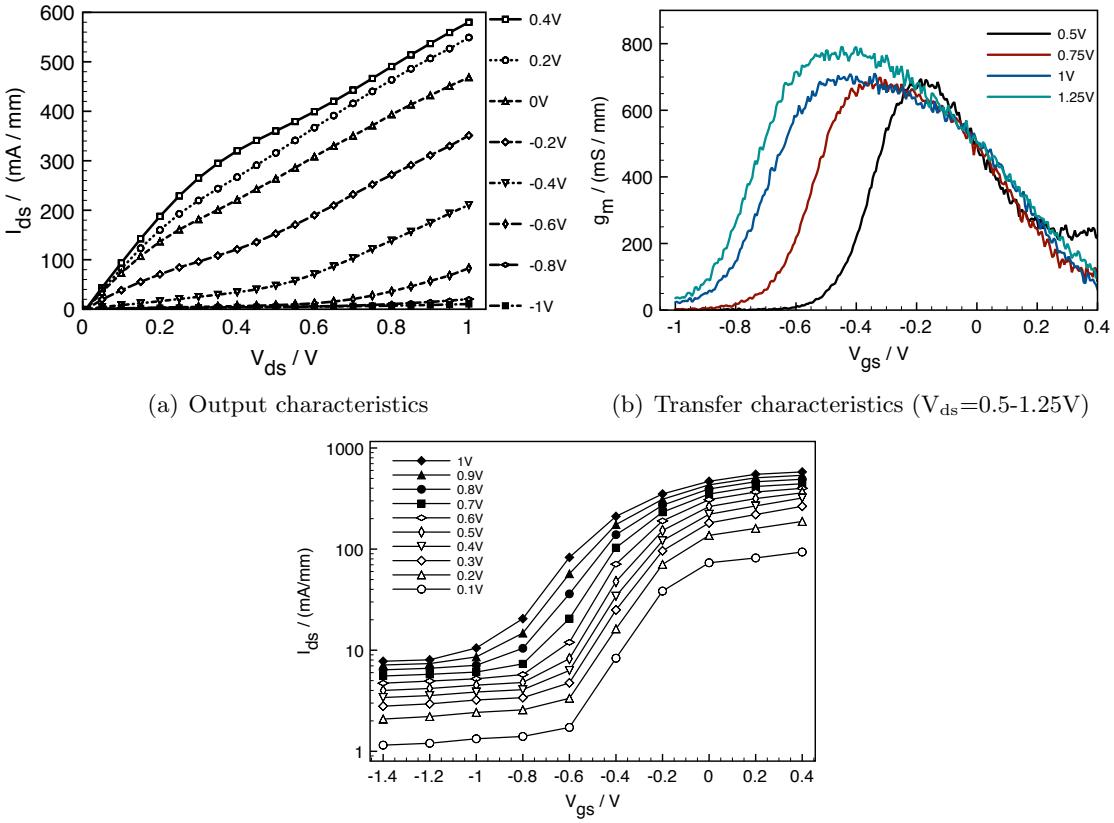
(a) Completed 25 μm device. (b) Gate region

Figure 7.27: Top-down SEM of completed 22 nm HEMTs.

The gate foot is entirely encapsulated, hence not visible in top-down imaging. Figure 7.27(b) shows the completed upper gate and ohmic contacts. The entire region between contacts and gate is encapsulated in silicon nitride.

Two device samples were processed, one on C216 and one on C217. None of the C217 devices exhibited gate control of the channel of any note. The C216 devices, however, pinched off excellently as shown in the logarithmic plot of Figure 7.28(c), with drain current swings of approximately two orders of magnitude across the bias range. The output and transfer characteristics for these devices are shown in Figure 7.28.

These initial results were encouraging, since the first devices fabricated were functional: indicative of robust processing during an intrinsically highly complex fabrication process



(a) Output characteristics

(b) Transfer characteristics (V_{ds}=0.5-1.25V)

(c) Logarithmic V_{ds}/V_{gs} plot for various drain voltages.

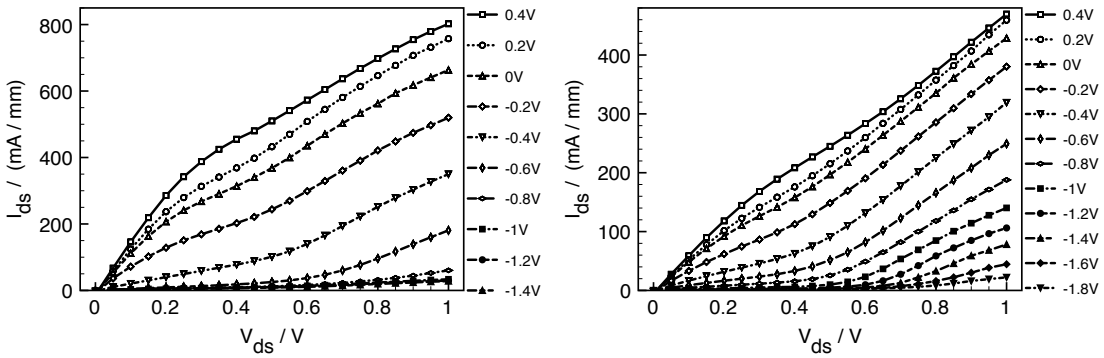
Figure 7.28: Output and transfer characteristics of initial 22 nm 100 μm-wide HEMTs.

flow. These results therefore confirmed the viability of the two-step gate process.

From the output characteristics of Figure 7.28(a), it is clear that for this 100 μm device, the V_{ds}/I_{ds} characteristics are far from ideal, with serious non-linearities in the saturation curves. Firstly, the devices do not truly saturate, though there is the onset of a pseudo-saturation regime at around V_{ds}=0.4 V. As such, the output conductance is extremely large. Output currents are also suppressed over prior devices fabricated on InP, such as the benchmark A1941, whose devices achieved saturation currents close to 1.4 A/mm [58].

These devices, however, achieved drain currents up to around 600 mA/mm, with threshold voltages of -1V. V_{gs}/Ids sweeps yielded peak transconductances of close to 800 mS/mm.

Device I-V characteristics, however, varied significantly with width, as evidenced by the V_{ds}/I_{ds} traces in Figure 7.29.



(a) 25 μm (b) 200 μm

Figure 7.29: Variation of 22 nm V_{ds}/I_{ds} characteristics with device width.

As is clear from the traces, there are major variations in both the peak output currents and the shape of the I-V curves with device width. The narrowest devices, at 25 μm wide, exhibit the highest normalised output currents, at up to 800 mA/mm at a drain bias of 1V with a positive applied gate voltage of 0.4 V. The narrow devices also have more ideal I-V characteristics - there is a more pronounced pseudo-saturation regime as compared to the 100 μm devices. Conversely, looking to the 200 μm devices, normalised output currents are considerably lower at a maximum of 470 mA/mm, with virtually no apparent expected linear/saturation regime and only small changes in conductance for each trace. One main feature of particular note is that whilst 25, 50 and 100 μm devices featured a uniform threshold voltage of approximately -1 V, the 200 μm devices required -1.8 V to reach comparable pinch-off.

As a consequence, it appeared that device width significantly affected transistor performance, beyond the usual high-frequency implications. In particular, wider devices suffered increased current suppression and spurious kink phenomena than narrower ones. A comparison of the various device widths with zero applied gate bias as the drain voltage was swept from 0-1 V is shown in Figure 7.30.

Comparing devices across the four cells on the sample also showed that two cells featured transistors which pinched off satisfactorily. The other two cells did not achieve pinch-off at higher applied drain voltages.

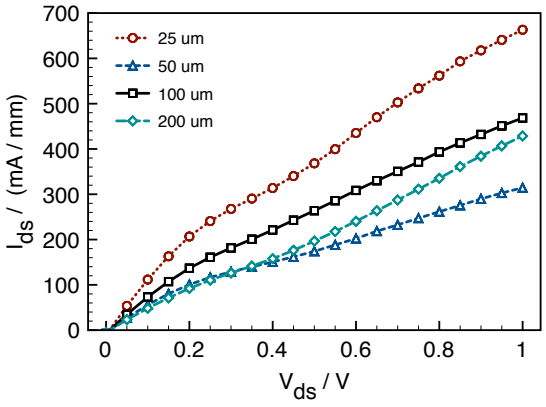


Figure 7.30: Comparison of zero gate bias V_{gs}/Ids traces for various device widths.

Transport data were also extracted from the samples following the complete device process flow. Van der Pauw measurements shown in Table 7.4 show that recessed figures of merit declined slightly during fabrication.

	Capped	Recessed
Sheet electron concentration / (cm^{-2})	6.3×10^{12}	1.1×10^{12}
Mobility / (($cm^{-2}v^{-1}s^{-1}$)	7740	10328
Sheet ρ / (Ω/sq)	128.1	540.4

Table 7.4: C216 van der Pauw transport properties following device processing.

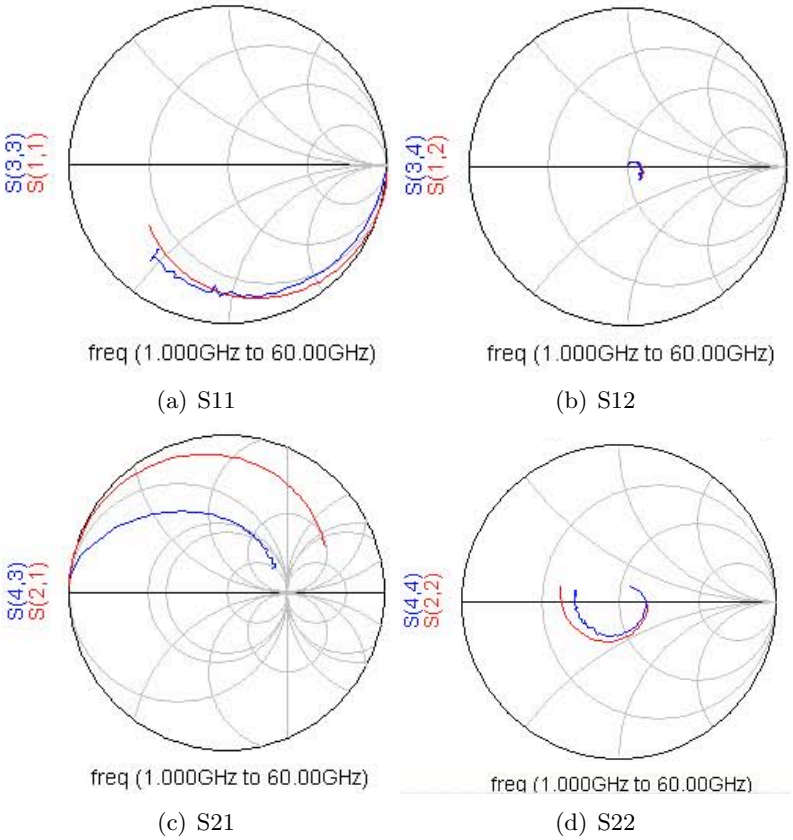
Whilst the capped measurements have changed negligibly, there have been significant changes to the recessed case, where mobility dropped by around 8%. Sheet electron concentration also dropped by 49%, leading to a doubling of recessed sheet resistivity. Since $I_{ds} = Wqnv$ (Equation 3.26) and both n and v have decreased as a consequence of the degraded channel transport, it is to be expected that resultant device drain currents will drop correspondingly.

The key points to note from the d.c. device data are therefore:

- Increasing deviation from expected FET characteristics with increasing device width.
- Suppressed drain currents for all devices with increasing suppression for wider devices.
- Threshold voltages are larger than expected (Approx -1V).

- Transconductances are much smaller than anticipated (0.8 S/mm maximum).
- Increased threshold voltage for wider devices.
- Large degree of variation between cells.
- C217 sample did not pinch off at all.

Given the encouraging, if non-ideal, d.c. characteristics of these initial devices, r.f. measurements were also taken from 0.1-60 GHz. In each case, data were extracted for the zero-bias, peak g_m and "cold" (forward-biased gate) methods. The raw data are shown as the red curves in Figure 7.31.



(a) S11 (b) S12

(c) S21 (d) S22

Figure 7.31: Smith chart plots of measured s-parameters and best model fit.

Analysis of the r.f. data was carried out in the Agilent Advanced Design System (ADS) software. The equivalent circuit described in Section 3.7 was then employed to model

the transistor. By building the model in ADS with included variable-length transmission lines on the gate and drain, the effects of the CPW transmission line structures that are realised at the bondpad level for each device can be isolated. The true performance of the device, inclusive of extrinsic parasitic parameters such as the axial resistances and parasitic capacitances, but exclusive of the transmission lines that do not comprise the device itself, can therefore be extracted.

By adjusting the model component values using optimisation techniques, a model which correctly fits the transistor characteristics can be extracted and used for comparison and projection purposes.

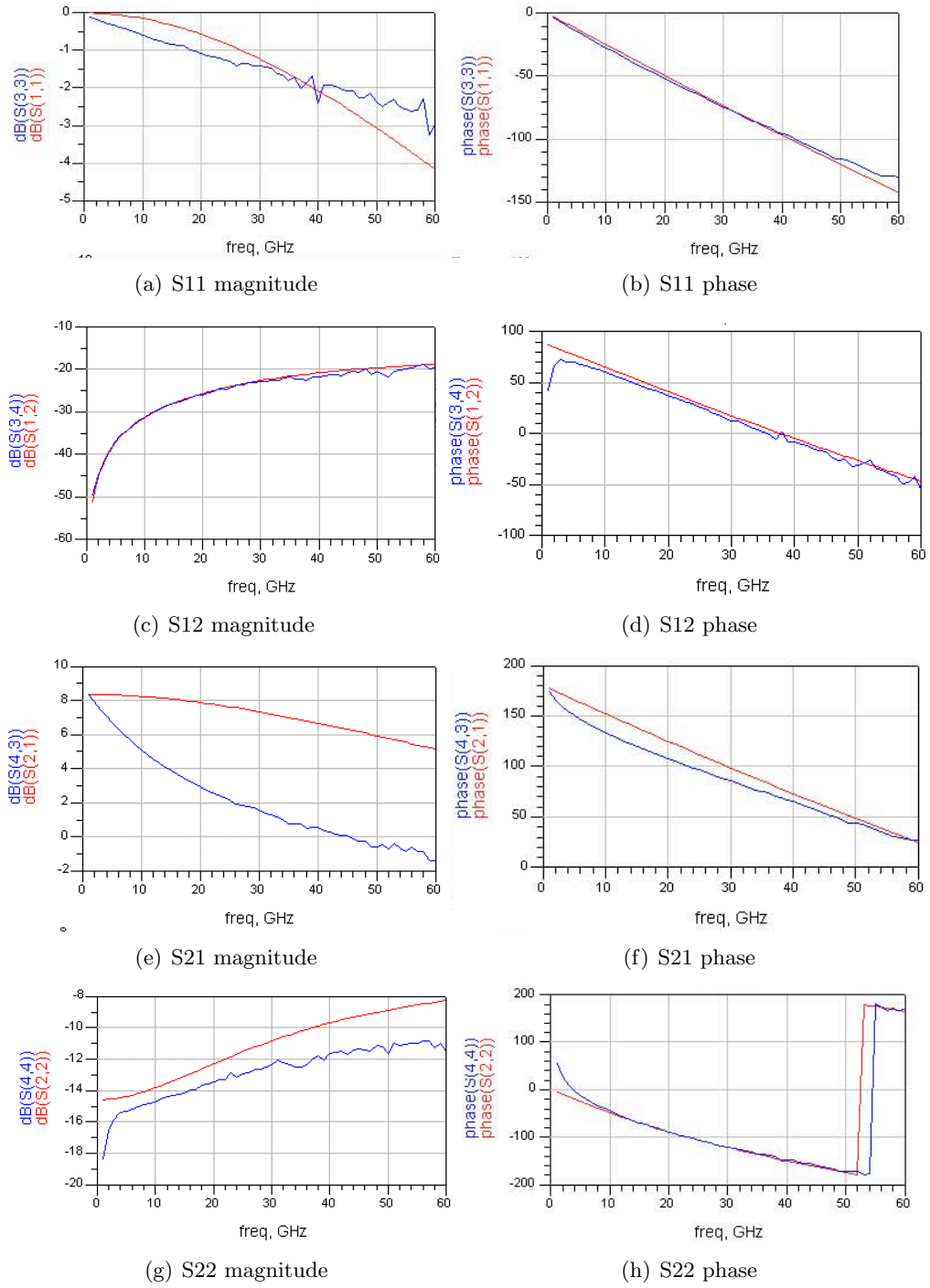
In the case of the s-parameter data for the 100 μm transistor shown in Figure 7.28, this proved a difficult process.

Whilst it was possible to fit the majority of the device characteristics using the standard model, it was impossible to fit them all satisfactorily by any method, predominantly due to the unusual dispersion of S21's magnitude. In general, S21 decreases following a slow exponential decay with frequency. In the case of the devices measured, however, S21 dropped rapidly with frequency, reaching 0 dB at only 45 GHz. The results of the best fit possible in ADS are shown as the blue curves in Figure 7.31. Additionally, the magnitude and phase of each plot are shown in Figure 7.32.

7.7.1 Discussion

The decay of S21 in this fashion is extremely unusual. At low frequencies, gain is as expected, with over 8 dB of gain evident at 20 GHz. As the frequency increases, however, gain decays exponentially rather than logarithmically. This may be indicative of transconductance dispersion, where the transconductance, hence the modulation efficiency, of the device varies with frequency, reducing output currents for a given gate voltage. Such effects have previously been noted at low frequencies [321–323], and in general are caused by the variable activation energy of surface states.

If the barrier were damaged in some fashion under the gate, then it is conceivable that traps underlying the gate might show a variable response time and activation energy and hence a frequency and bias dependence. The damaged surface might also exhibit a perturbative effect on the carriers in the channel if the surface pinning were affected; in effect, altering the surface potential. The unusual device output characteristics might hence be explicable by the same theory. The general suppression in drain current and



(a) S11 magnitude (b) S11 phase

(c) S12 magnitude (d) S12 phase

(e) S21 magnitude (f) S21 phase

(g) S22 magnitude (h) S22 phase

Figure 7.32: Magnitude and phase of measured and modelled s-parameters from the first run of devices. It is clear the S21 fit is poor.

relatively large threshold voltages might also be inter-related to surface damage.

A damage-related theory, however, does not explain all the unusual characteristics. The lack of pinch-off from one sample and variations between cells are likely lithographically and process-dependent. In particular, it seemed likely that there might be regions masked or unmasked by imperfect lithographic alignment: a variable between cells and between samples.

A further complication was that the recess was designed as deliberately large for these devices to minimise the effects of gate-recess misalignments. A larger InAlAs surface was therefore exposed than would otherwise be the norm, with potential exacerbations to any surface abnormalities resulting from the device processes.

Cross-sectional SEM images were attempted in a Hitachi S900 SEM to verify both the successful completion of the device processes and to provide any clues to surface damage or oxidation. The samples, however, proved extremely difficult to image, for several reasons. Firstly, the critical geometries involved are extremely small and intrinsically difficult to image, requiring the highest possible resolution of the microscope and close to perfect control of beam focus and stigmation, particularly when examining the interfaces and surfaces around the gate. Secondly, there are major issues in imaging the silicon nitride surrounding the gate. There is an intrinsic stress in the deposited film, which tends to cause it to cleave non-uniformly. The surface to be imaged therefore tends to become badly damaged during sample preparation. As a further complication, despite sputter coating, the silicon nitride, being an excellent insulator, charges badly during imaging, particularly at the high accelerating voltages desirable for high resolution. As a consequence of these effects, it was clear a more suitable analysis technique would be required for future device inspection.

It is additionally worth noting that at the time of fabrication, no devices had been reported with such short gate lengths, though the 562 GHz Fujitsu device was claimed to be 25 nm. As a consequence, the short gate length effects were essentially unknown, and a great deal of the performance abnormalities may have arisen from issues of transport or aspect ratio.

The decision was therefore taken to repeat the device process run with two major changes; reduction of the recess length and realisation of devices with a range of gate lengths to isolate and analyse any gate-length-dependent effects.

7.8 Second-generation C216 devices : gate length variation

50 nm devices are commonly fabricated on substrates similar in terms of composition and layer structure to C216 and C217 at Glasgow. For comparison purposes, therefore, it was decided to fabricate a spread of gate lengths from the minimum possible using the new gate process to the usual 50 nm, allowing materials issues to be decoupled from those relating to the gate process itself. To realise devices with gate length varying uniformly from the minimum 22 nm of the previous batch of devices to a nominal 50 nm, doses were extracted for the ZEP520A / SF_6/N_2 processes shown in Figure 7.13 using a simple linear fit to the previous exposed test samples.

The required doses for a spread of gate lengths are shown in Table 7.5. Since the etch process yielded trenches uniformly smaller than the resist mask, doses were chosen to yield features at approximately 5 nm gate length steps in addition to 22 nm: 30, 35, 40, 45 and 50 nm, though in most cases, slightly smaller features were expected.

Dose / μCcm^{-2}	Resist linewidth / nm	Si_xN_y linewidth / nm
1700	23.1	22.1
3000	31.3	29.8
4100	36.1	34.5
5100	40	38.2
5765	44.3	41.7
7350	50	48.5

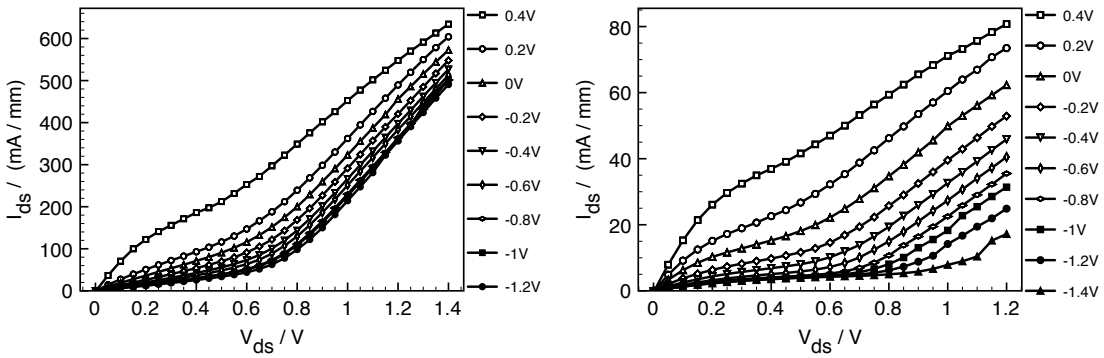
Table 7.5: Doses assigned for approximately 22-50 nm device realisation.

Devices of various gate lengths were then fabricated using processes broadly identical to those of the first-generation devices, both on C216, and on A1940, the lattice-matched equivalent to A1941. The reasoning for the use of apparently inferior material was to eliminate the new layer structures as the cause of the undesirable device characteristics seen on C216/7, since A1940 had previously produced excellent devices.

One change, however, was the realisation of a shorter recess trench, whilst maintaining the double-recess profile.

Instead of the previous 12s / 5s / 60s succinic acid / orthophosphoric acid / succinic acid etch times, 12s / 5s / 10s etch times were used, reducing the recess dimensions considerably from 150 nm to around 80 nm, shortening the recess trench with the hope of reducing any surface effects, but increasing the device alignment requirements. As

previously, the recessed surface was briefly treated with hydrofluoric acid to deoxidise the surface prior to silicon nitride deposition.



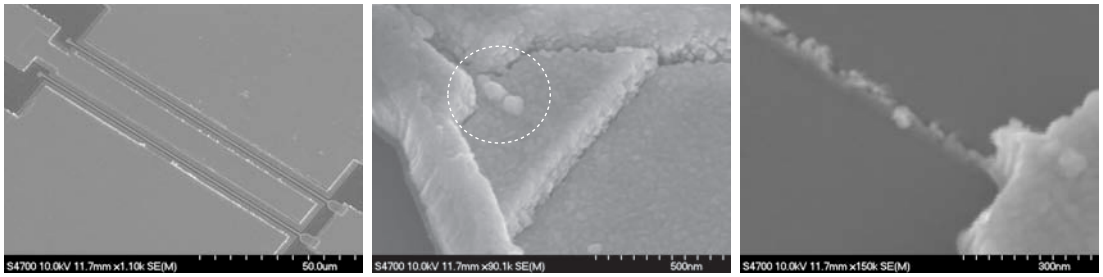
(a) A1940 (2×50 μm, L_g = 45 nm) (b) C216 (2×50 μm, L_g = 50 nm)

Figure 7.33: Typical Ids/(V_{ds}, V_{gs}) characteristics of second-run devices.

Initial devices fabricated with the revised process flow were disappointing; none of the devices on either sample pinched off (Figure 7.33). Of particular note was the fact that C216 devices, though apparently approaching the pinch-off condition to a greater degree than the A1940 devices, had drain currents almost an order of magnitude smaller. In addition, both sets of devices still suffered from similar kinks in the I-V characteristics to the initial devices, despite the shorter recess profile. As a consequence, the recess process alone was insufficient to explain the unusual device characteristics seen in the first-generation devices.

The devices were examined closely by SEM to inspect for any issues to explain the poor pinch-off characteristics.

One spurious feature was noteworthy from the images: the presence of an unusual raised area under the gate feed, annotated in Figure 7.34(b). Whilst the central area of the upper gate shows a dip in the metallisation at the "foot" region, since the evaporated metal coating is non-conformal, in the feed area, where the feed overlaps the first gate metal below, the true topography of the second gate metallisation is evident. It would seem that there is an additional raised area in the region of the lower gate metal, caused either by non-uniform metal filling extending above the silicon nitride trench during either the first gate metal evaporation, or by effects of compounded metallic grain size. As is evident from Figure 7.34(c), taken at a site where the upper gate metal lifted off poorly across the gate width, but remained in the feeds, the upper gate metallisation is



(a) SEM overview

(b) Gate feed at mesa overlap. The raised area does not extend to the mesa edge.

(c) Failed lift-off showing origin of raised area.

Figure 7.34: Completed devices and gate feeds showing misalignment at mesa edge.

connected to the lower gate foot, and in regions where the lift-off was unsuccessful, the metal extends above the etched silicon nitride trench.

As a consequence, it was concluded that the first gate metal did not extend as far as the edge of the mesa.

In a conventional HEMT process flow, the gate feeds are realised in the same lithographic exposure as the gate foot, the recess etching using the gate resist as an etch mask for the wet etch. The area below the feed is thus recessed at the same time, ensuring complete etching of the cap along the whole gate width. In the case of the new process flow, however, the gate feeds are realised with the upper gate only, with the cap recessing taking place before even the first gate metal is deposited to ensure encapsulation in silicon nitride. The etch mask for the recessing for the initial devices was designed to be identical to the area exposed for the first gate metal: a single pixel line of identical width.

As a consequence, if the first gate metal failed to extend over the mesa edge, and assuming repeatable alignment accuracy, the surface between the end of the first gate metal and the edge of the mesa would have the cap intact. From the images of Figure 7.34, this region appears to be around 200 nm in width after isolation etching. Accordingly, the cap remained in place for the devices examined in the SEM at the edge of the mesa, leading to a sizeable cap region, highly populated by electrons in addition to the channel population. As a consequence, the action of the gate in increasing the surface potential with increased negative bias will have greatly reduced effect on the channel population in this region, leading to increased source-drain currents. In the extreme case, current

remains for any gate bias.

This explanation describes the scenario for both sets of initial devices and suggests reasoning for the variation in threshold voltage and pinch-off between devices on the same sample and between samples. Slight variations in alignment accuracy would lead to increased or decreased overlap between the two gate levels, the recess and the mesa edge. In addition, variations in etch conditions would have affected the lateral etch rate and hence undercut of the designed isolation edge (Section 4.5.1), leading to differences in the area of the remaining unetched cap.

Two actions were taken as a result of this realisation:

- A scheme was attempted to etch through the upper gate metal and underlying silicon nitride in the capped region in order to selectively recess the area in the fabricated devices, verifying the theory. An argon sputtering dry etch approach was adopted to etch the gold of the upper gate metal with sufficient control. Unfortunately, the sputter process was not able to provide the etch profile required, with significant re-deposition of the gold around the etched areas and the devices were abandoned.
- Alterations were made to the GDS layer design of the devices to allow for misalignment and process variations. The lower gate width was extended on both sides of the mesa edge under the upper gate, whilst a small "notched" section was added to the recess level at the feed end of the gate to allow for additional misalignment.

The unusual I-V characteristics of the devices, however, remained unexplained by this finding. Confirmation of the unusual results on A1940 allowed the new wafer designs to be eliminated as the cause. Since identical process steps were used as for the previous batch of devices, it seemed reasonable to conclude there was a step inducing damage to the underlying materials.

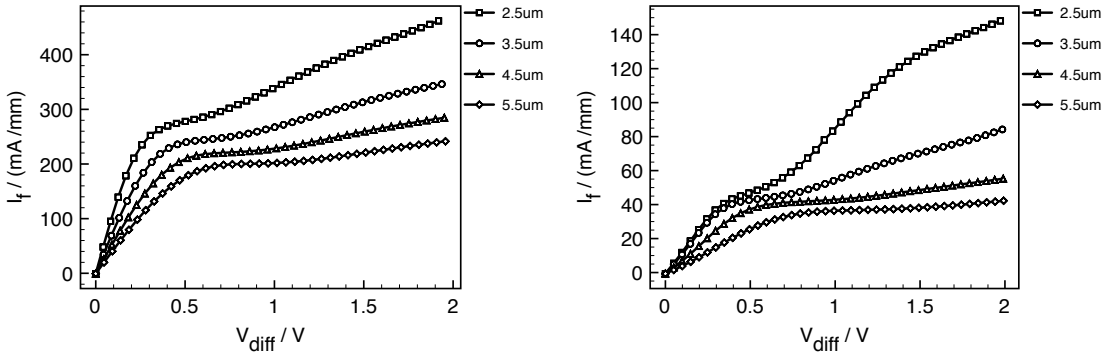
TLM techniques were also used to extract the contact resistances of the two samples following complete device processing. In contrast to the first-generation devices, the cap etch regions for the TLM and van der Pauw sites on these samples was defined in the same lithographic step as the gate recess for the devices. As a consequence, their recessed characteristics should show similar characteristics to the devices, having undergone identical processing.

	A1940		C216	
	Capped	Recessed	Capped	Recessed
Contact resistance / (Ω.mm)	0.045	0.051	0.12	0.091
Sheet ρ / (Ω/sq)	101.2	462.7	118.6	4967.2

Table 7.6: C216 and A1940 TLM data for completed device samples after complete device processing.

It is evident from the extracted resistivity data that there are significant processing problems; for both samples, extracted capped contact resistances and sheet resistivities are comparable with those of Table 7.3. In the recessed case, however, whilst the contact resistances appear to be of the same order of magnitude as expected, the sheet resistivities have increased by 66% in the case of A1940 and 1320% in the case of the C216 sample. Both increases are massive, since slight transport changes directly impact on drain currents and electron velocity, though the C216 case is particularly severe.

One explanation for the differences between measured A1940 and C216 TLM data might be the reduced barrier thickness of C216 (10 nm instead of 15 nm for A1940) as well as the reduced channel thickness (also 10 nm as opposed to 15 nm). Both differences would exacerbate the effect of any damage to the semiconductor surface: changes to the surface potential would cause increased perturbation to electrons closer to the surface: precisely the result of a thinner barrier and channel.



(a) A1940 (150 μm TLM width)

(b) C216 (150 μm TLM width)

Figure 7.35: I-V characteristics of varying TLM gaps. It is clear that C216 exhibits greatly suppressed currents and additional kinks in its profile in contrast to A1940.

In addition, examination of the actual four-probe I-V traces from the TLM data indicates a familiar transfer characteristic: identical profiles to those seen in the device character-

istics of both fabrication runs, as is evident from Figures 7.33 and 7.35. C216 shows vastly reduced drain currents and additional non-uniformity in its transfer characteristic. It is therefore clear that the device characteristics observed have their origins in the same effects that significantly increased the TLM sheet resistivity.

With this in mind, it was decided to conduct a thorough investigation into changes to a recessed InAlAs surface on C216 during various process steps. It was also felt that increased process control over the recessing process was required.

7.9 InAlAs surface processing

The InAlAs surface was exposed to several different chemistries during the process flows of the initial devices, based on prior conventional process flows. Whilst each may have been convenient for its original purpose, the combination may not be suitable to the new process. It was therefore important to separately investigate each of the following possible causes of damage to the exposed InAlAs:

1. Recess etching
2. Post-recess clean (conventionally using H.F. acid)
3. Silicon nitride deposition
4. RIE and other plasma etch processes
5. Pre-metallisation de-oxidation
6. Metal evaporation

Each of the above processes involves an exposed InAlAs surface and may therefore result in induced damage. Several of the steps, however, had previously been eliminated as causes of damage: silicon nitride deposition effects were discussed in Section 7.4.2, whilst identical metal evaporation processes were used as for all prior devices. The remaining potential causes were the post-recess clean, RIE silicon nitride etch and de-oxidation steps.

7.9.1 Silicon nitride overetch

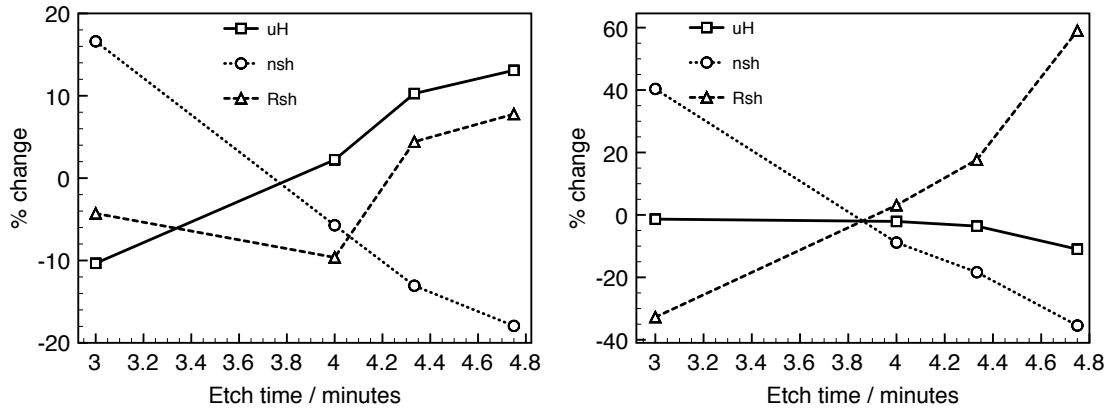
Exposure of the InAlAs surface to the SF_6/N_2 RIE used during etching of the silicon nitride is likely to result in some degradation of the transport properties, despite the low-damage nature of the etch process. During the gate etch, it is known that the device regions are not subject to excess exposure to the process gases due to the small area of the etch regions. This effect is the basis for the additional over-etching time required to complete the etch of the short gate foot.

It is expected that the etch process itself should not damage the underlying material as long as the surface is not in contact with the plasma; in effect, until after the etch process has terminated. As a consequence, it is expected that damage would only occur during excessive over-etching. Additional over-etching, however, could occur accidentally, or if large areas were etched simultaneously with small areas.

To confirm this, recessed van der Pauw and TLM structures were realised as previously on C216 and expose to various etch times in SF_6/N_2. The samples were unmasked and hence a "blanket" etch process was used to etch the entire sample. Shorter etch times than would be required for smaller structures would therefore be expected, given the increased probability of reactant diffusion to the surface to be etched. Reflectometry had previously showed the end point of the etch process to occur at approximately 3.5 minutes for this process, hence etch times were chosen at intervals after this point. In addition, one 3 minute etch was carried out. The samples were cleaved from a single larger sample to ensure parallel processing and a control sample kept unprocessed.

The various figures of merit were then measured using the van der Pauw technique and changes for each etch time noted with respect to the control sample. The changes for both capped and recessed cases are shown in Figure 7.36.

As expected, for the under-etched case, the same metrics apply as those seen in Figure 7.4.2: in the capped case, electron density is increased but mobility reduced, whilst in the recessed case, mobility remains virtually unchanged whilst electron density increases markedly. As the etch time is increased above the threshold for complete film removal, sheet electron density drops rapidly in both capped and recessed cases. Interestingly, in the capped case, mobility increases with over-etch, whilst it deteriorates slowly in the recessed case. In both cases, the net effect is to increase sheet resistivity for large over-etch times. Also as expected, the recessed case exhibits far greater sensitivity to etch-induced damage: clear from the scales on the data. In the extreme over-etch time of



(a) Capped (b) Recessed

Figure 7.36: Changes in mobility, sheet electron density and resistivity with varying etch time.

4m 45s, sheet resistivity increased by nearly 60%, whilst sheet electron density dropped by over 35% in the recessed case. Due to the high initial cap electron density, this contrast may be unsurprising.

These low-field results are as expected, indicating increasing damage for increasing over-etch times. The strange kinks in the device I-V characteristics, however, imply a field-dependence to the damage phenomena. As a consequence, the etched TLM sites were also measured using usual four-probe TLM techniques. The results are shown in Figure 7.37.

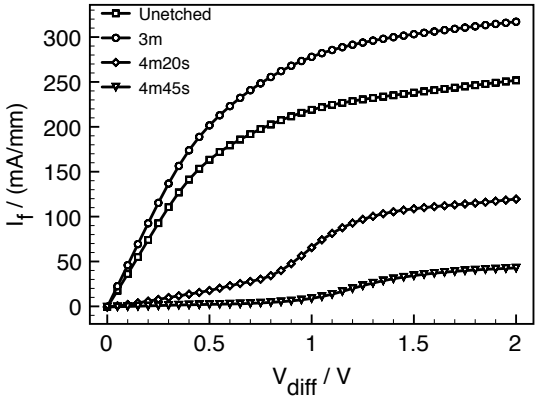


Figure 7.37: I-V characteristics of 3.5 μm TLM gap (150 μm wide) using samples from Figure 7.36. Unfortunately, the 4 m sample was damaged and unuseable.

The results are interesting: in the under-etched case, current enhancement occurs, as might be expected considering the constant mobility and sheet electron density with silicon nitride in place for recessed material. This is also a confirmation that, though mobility tends to drop in capped regions, the net effect across a region with both recessed and capped areas is enhancement. This relationship may, however, be variable if the recessed region is much smaller than the capped region.

Unfortunately the 4 m etch sample was damaged at the 3.5 μm site and no reasonable comparison can be made. For longer etch times, however, it is clear that current indeed becomes suppressed, with the introduction of spurious kinks in the characteristics. It is worth noting that in the extreme over-etch case, current drops by over two thirds, comparable to the 60% higher resistivity seen in the Hall measurements. The current suppression is, however, much greater in low-bias regions, before the kink in the characteristics.

It therefore seems reasonable to conclude that long extraneous over-etching, as expected, induces damage which could explain the kinks in the device characteristics.

In the device case, however, the InAlAs is necessarily only briefly exposed to the etch chemistry. Any additional over-etch would significantly enlarge the silicon nitride profile as is evident from Figure 7.12. Consequently, though plasma damage could clearly explain the unusual behaviour of the devices, this is extremely unlikely.

A further point of particular note, however, is the role that over-etching may play in the formation of low-resistance ohmic contacts. As shown in Figure 7.36(a), capped sheet electron density drops considerably with a corresponding increase in resistivity with increasing SF_6 over-etch. Gross over-etching of the ohmic contacts will markedly impair transport into the channel, degrading contact resistance. It therefore becomes important to optimise the ohmic etch process, though the requirement for precision is considerably relaxed over the gate etch as a consequence of the reduced sensitivity of the InGaAs surface in comparison to the exposed barrier. As part of this work, a four minute RIE was found to optimally etch the large-area ohmic contacts.

7.9.2 Post-recessing surface treatments

Whilst plasma-induced damage is unlikely to explain the variable current suppression, it is now clear that a surface damage mechanism can explain the observed effects. There are therefore three further variables which may induce additional damage during the device

process flow previously adopted:

- Post-recess acid clean.
- Exposure to an SF_6 chemistry. Though plasma damage is unlikely to be a dominant factor, the presence of fluorine has previously been shown to passivate dopants. Exposure to a fluorine chemistry and subsequent thermal cycling, enabling diffusion, may yet be relevant [324].
- Pre-metallisation de-oxidation.

The points are inter-related, since the first and third involve the same acids and bases for the removal of oxides at various stages of the process, whilst the etch chemistry may interplay with additional surface traps induced at an earlier stage. A processing matrix was therefore devised involving multiple different substrates to examine the effects of various different chemistries on the four-probe I-V characteristics of recessed TLM gaps, fabricated identically to previous experiments.

It was suspected that either the hydrofluoric acid (HF) or SF_6 exposure were damaging the InAlAs surface in some way, possibly by modification of the surface states in the recessed region. One possible method for the partial long-term passivation of surface states, suggested in myriad publications over many years, is sulphidation. A sulphidation process generally makes use of a sulphur-containing compound to deliver sulphur to the surface, where it is adsorbed and thought to bond with some surface states, reducing their reactance. Though never conclusively proved, sulphidation processes have shown considerable promise if correctly prepared in various publications [325–331].

In addition, whilst the influence of a plasma was unlikely, gaseous SF_6 chemistries might still have a deleterious effect on drive currents as a consequence of donor passivation [324, 332–335]. In particular, fluorine exposure followed by thermal cycling have been shown to have significance in passivating dopants.

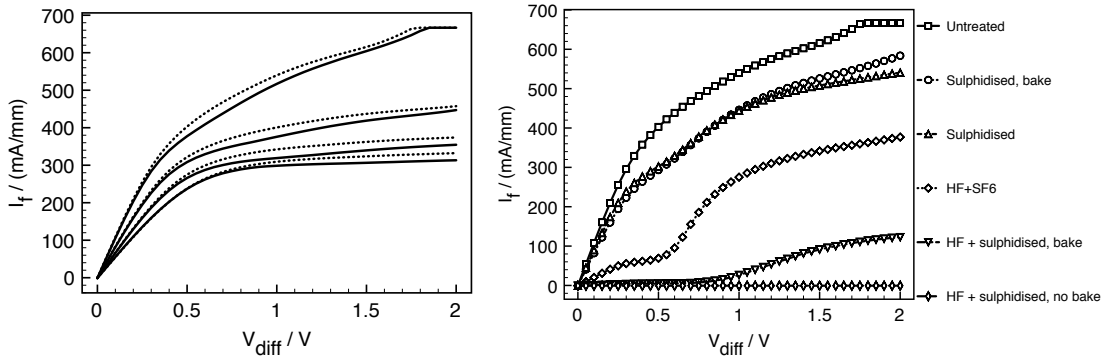
As a consequence, it was decided to incorporate three major process variables into a test matrix on C216 to analyse the effects of the various processes both separately and in combination:

- HF de-oxidation. A 100:1 HF:H_2O mixture was previously used following a recess

etch and prior to silicon nitride deposition to remove any residual etch products from the surface.

- Exposure to SF_6. The gas was delivered to the surface for 30 s using the System-100 etch tool without striking a plasma. Any changes are therefore related purely to the chemical interactions at the given temperature and pressure.
- Sulphidation treatment. A dilute 10:1 ammonium sulphide : H_2O mixture was used. Some samples were also "annealed" at 180°C for two hours following sulphidation to allow increased diffusion and removal of excess sulphur.

A summary of the results of the process matrix is shown in Figure 7.38, which compares the effects on a 2.5 μm recessed TLM site. The full I-V results for various gap dimensions for each treatment are available in Appendix B.



(a) SF_6 exposure effects on recessed TLM (b) Comparison of various combinatorial treatments on a 2.5 μm recessed TLM. Dashed traces are the pre-treatment curves.

Figure 7.38: Effects of various treatments on TLM I-V characteristics.

It is clear from Figure 7.38(a) that exposure to SF_6 alone has negligible effect on the recessed surface, yielding I-V characteristics virtually identical to the pre-treatment values. The use of HF, however, yielded dramatically different results, with massive current suppression and non-linearity appearing immediately. The kink profile is additionally very similar to that seen in early devices. In conjunction with SF_6, further increases in suppression occur. Sulphidation also appears to produce some damage in conjunction with SF_6, which may be unsurprising given its surface cleaning action. Baking the samples produces negligible change.

The most drastic results are the combination of an HF surface clean and a sulphidation treatment, the accepted "best" sulphidation technique for surface cleaning and passivation [328]. In conjunction with SF_6, the process is adequate to eliminate all current flow. The addition of a baking step does yield some recovery of the current under large electric fields.

The majority of the process damage appears to be due to the use of HF, as opposed to the plasma processing. Given the severity of the damage from the use of HF, various alternative de-oxidation agents were tested by the same metric: measurement of recessed TLM structures. Various concentrations of buffered and unbuffered HF were tested, as well as sulphuric, orthophosphoric and hydrochloric acids and ammonium hydroxide. The results are shown in comparison for a 2.5 μm TLM gap in Figure 7.39. Again, the full results are available in Appendix B.

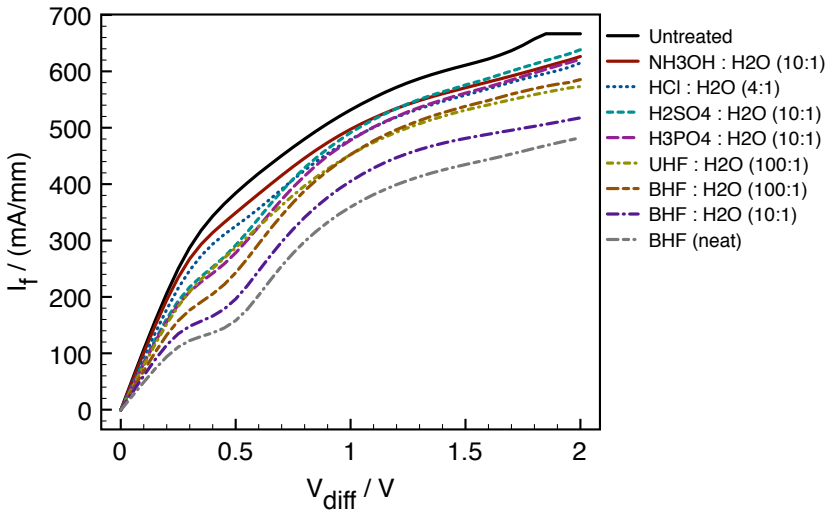
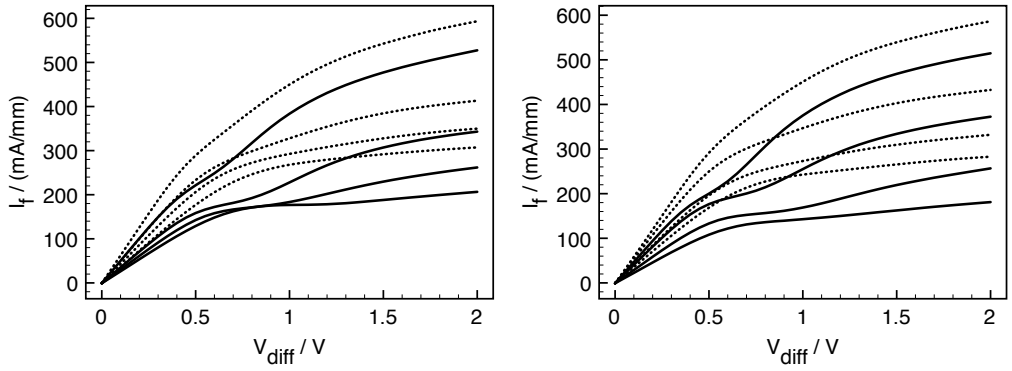


Figure 7.39: Comparison of various de-oxidation treatments on a 2.5 μm recessed TLM.

From the various extracted data, it is clear that all the acid-based etchants inflict some damage to the surface, with HF giving particularly poor results. It is also particularly clear that it is truly HF, not its buffering solution, that causes the damage, as damage is evident after exposure to both buffered and unbuffered solutions of equivalent concentration. Damage clearly also decreases with decreasing concentration, since more dilute HF concentrations yield less pronounced effects.

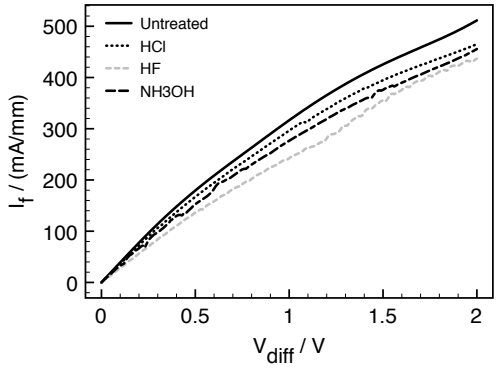
All the remaining acids are less damaging, but do substantially reduce the drive currents. Interestingly, ammonium hydroxide, the only non-acid oxide etchant, apparently does not produce the same damage, with I-V characteristics closely following the pre-treatment traces. Hydrochloric acid is also less damaging, but does produce further kinks in the I-V characteristics.

In order to separate any potential damage to the unmasked ohmic contacts, a set of experiments was also run on C217 samples where the recess resist remained in place over the test sites, protecting the contacts from the treatment solutions. As is clear from Figure 7.40, damage still occurs in the masked case and is virtually identical to the unmasked sample results. Damage to the contacts would therefore not appear to be the dominant damage mechanism in previously measured samples.



(a) BHF:H_2O (1:10) on recessed TLM sites. Dashed traces are the pre-treatment curves.

(b) HCl:H_2O (1:4) on recessed TLM sites. Dashed traces are the pre-treatment curves.



(c) Effects on 2.5 μm TLM where the recess length is 100 nm.

Figure 7.40: Effects of various treatments on TLM I-V characteristics where the contacts are masked.

A set of TLMs was also fabricated where the recess was defined using a short etch of approximately 100 nm and the resist kept in place during the post-recess treatments of hydrochloric and hydrofluoric acids and ammonium hydroxide. As was apparent from the results of Figure 7.40(c), though the various treatments produce a spectrum of effects on the transfer characteristics, as previously, the apparent damage is far less significant, with more minor deviations from the untreated characteristics arising, as is clear from a comparisons between the masked HCl case of Figure 7.40(b) and that of Figure 7.40(c). As with previous tests, both hydrochloric acid and ammonium hydroxide produce the least deviations, whilst HF continues to produce spurious kinks and current suppression.

The reasoning for the reduced damage between the bulk and short recess cases is unclear, but may relate to diffusion of the reactants on the surface. Several things are clear from this series of experiments: the use of aggressive acids on this material system appears to be largely responsible for the damage effects apparent in devices; ammonium hydroxide, the only non-acid treatment agent, yields the least damage of all treatments considered; treatments on an unmasked surface greatly exacerbate the effect and larger recessed areas are more susceptible to damage during treatment.

As a consequence, all non-essential acid treatments were removed from the device process flow. Previously, HF was used prior to silicon nitride deposition after recessing to remove any contaminant that may remain: carried out on the bulk surface with the resist mask removed. Hydrochloric acid was then used for additional de-oxidation immediately prior to gate metallisation. To avoid the damage caused by the use of acid-based de-oxidation or cleaning steps, the post-recessing clean was removed prior to silicon nitride deposition. The process flow is hence increasingly similar to a more traditional process flow, with only a single deoxidation/cleaning step in the recessed region.

In order to minimise potential damage from the use of orthophosphoric acid in the double recess process described in Section 7.7, only the first step of the recessing was performed, using an intermediate time to those used previously of 25 s, forming a single-step recess process. In order to reduce the potential of shorting the gate to the cap in the event of one of the longer gate lengths approaching the edges of the expected 60 nm recess trench, a short succinic acid step was also used after the etching of the silicon nitride during the formation of the gate foot. Areas in which the cap has been removed, as anticipated, should be unaffected by the selective etch. Areas in which any traces of cap remain should be etched, leaving no trace of cap under the gate.

The pre-gate metallisation de-oxidation was then replaced by an ammonium hydroxide dip in accordance with the findings of Section 7.9.2.

It was anticipated that having identified and eliminated several potential sources of damage in the process flow, improved device performance might result.

7.10 Third-generation device results : gate length variation

Taking the results of the surface treatment experiments into account, devices were again fabricated at a range of gate lengths from 22-50 nm, using the gate dose and etch details of Table 7.5 on c216. The revised process flow was:

1. Markers
2. Recess - with no post-etch treatment, including revised recess/mesa geometry from previous devices. A 20 s single succinic acid recess step was used.
3. Silicon nitride deposition
4. Marker protection
5. Gate 1 - including ammonium hydroxide de-oxidation.
6. Isolation
7. Gate 2
8. Ohmic
9. Bondpad

As previously, devices were fabricated at a range of widths from 25-200 μm using the revised gate and recess layout to ensure continuity of the recess over the complete device width.

The device yield was excellent in one cell in particular, in contrast with earlier devices. D.c. device data were extracted systematically, using a semi-automatic probe station. As is evident from the I-V characteristics of Figures 7.41 - 7.44(c), the measured device characteristics were free from kink irregularities, with the exhibited output currents much larger than those of any previously-fabricated devices. It is therefore reasonable to assume

the conclusions of the surface processing experiments of Section 7.9 in conjunction with the modified recess layout explain the unusual characteristics seen in early devices.

Yield was extremely variable between processing cells. Some cells featured 100% functioning devices of a given gate width, whilst others yielded 0%.

The functioning devices yielded d.c. performance far beyond that of prior devices, with peak drain currents well over 1 A/mm measured over the range of devices fabricated. As shown in Figure 7.41, the 22 nm devices pinched off satisfactorily with a threshold voltage of around -1.2 V and a peak saturation current of over 1.2 A/mm. A peak transconductance of 1 S/mm was also achieved.

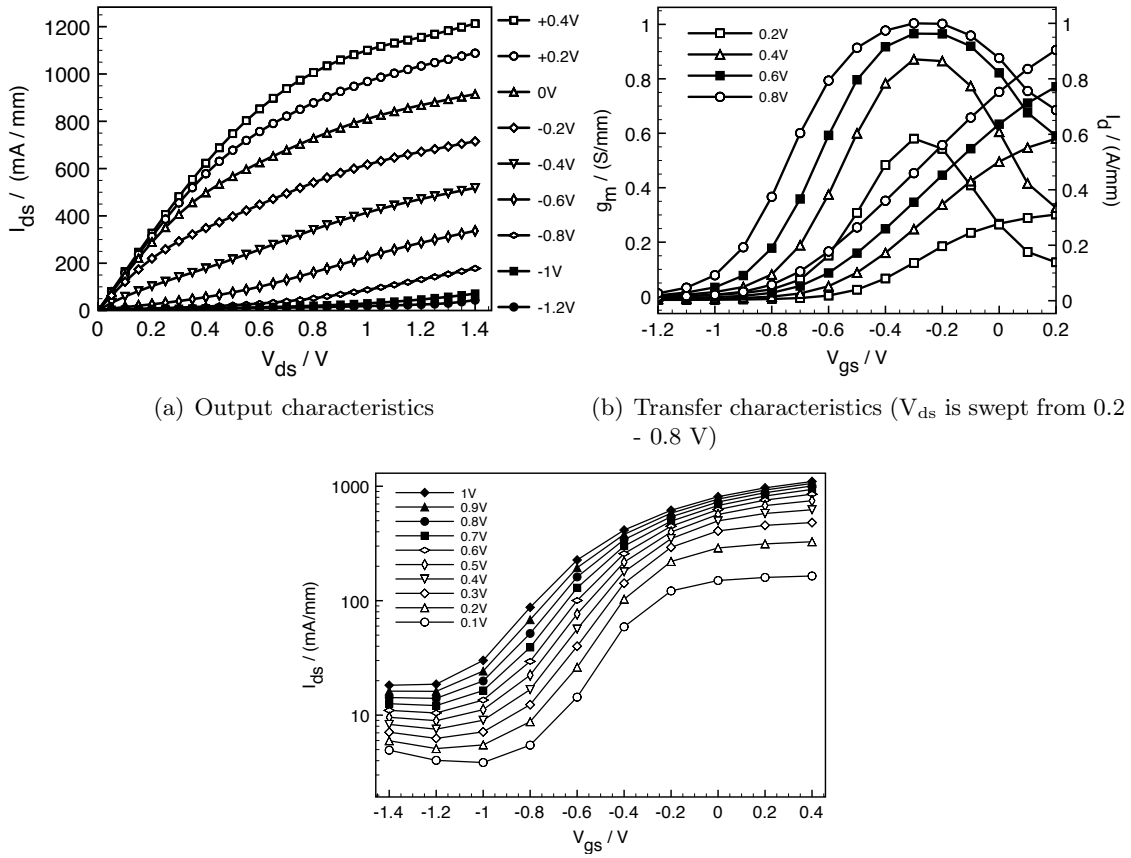
In comparison to the previous devices of Figure 7.28(c), off-state currents were uniformly slightly increased by approximately 3 mA/mm for all drain voltages, which may be a consequence of the change in surface treatments. The ratio of on and off currents, however, has increased from previous devices as a consequence of increased drive current.

Interestingly, longer devices in general exhibited increased gate control, with 50 nm gate length devices achieving peak transconductances up to 1.6 S/mm, as shown in Figure 7.42. These devices retained kink-free normalised drain currents of over 1 A/mm.

This trend was exhibited across the sample, indicating that the effect was not limited to one cell. Drain current, however, remained approximately constant with gate length: for each fabricated cell, the shortest functional gate length device exhibited a drain current on average 97 % of the longest working device. Its transconductance, however, was less than 80 % of the longest device on average. It would therefore appear that whilst the full range of gate lengths was fabricated with high yield, the lack of material scaling may genuinely be affecting device performance.

As gate length decreases, an increased average electron velocity can be expected as a consequence of the non-equilibrium transport phenomena of Section 3.8.2. Channel transport may therefore be enhanced (or at least unharmed) by the reduction of the gate length. If the aspect ratio of defined gate length to gate-channel separation drops, however, transconductance is expected to decrease (Equation 3.36), assuming all other parameters constant.

By retaining the gate-channel separation and reducing the gate length, the effect on the capacitive coupling of the gate to the 2DEG is effectively reduced, decreasing its modulation efficiency. In this case, the gate length has been reduced by 44 %. C216



(a) Output characteristics

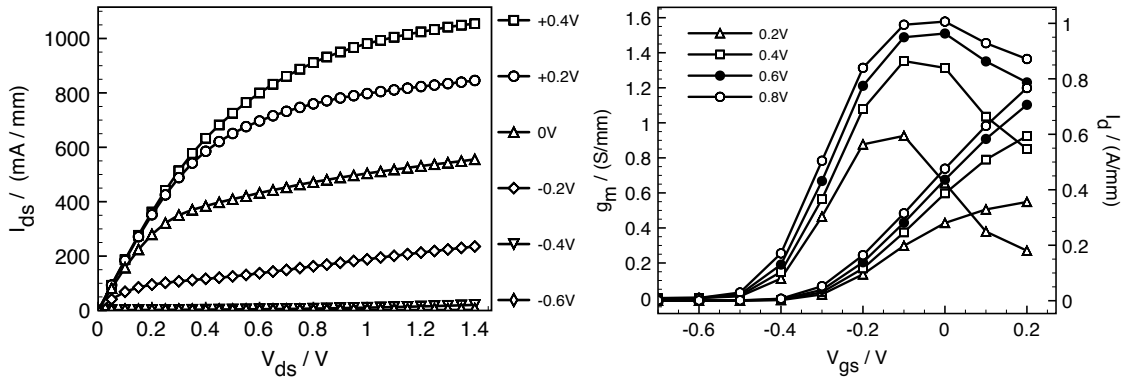
(b) Transfer characteristics (V_{ds} is swept from 0.2 - 0.8 V)

(c) Logarithmic V_{ds}/V_{gs} plot for various drain voltages.

Figure 7.41: Output and transfer characteristics of 3rd-generation 22 nm 50 μm-wide HEMTs.

has a gate-channel separation of 15 nm. Maintaining the aspect ratio would imply a gate-channel separation of 7 nm. Although one cannot extrapolate the transconductance as a percentage of that of the original aspect ratio, this does explain the drop. The comparative performance of the short gate length devices with respect to that of the 50 nm devices is shown in Figure 7.43. From this comparison, the general trend of increasing drain current but decreasing transconductance with decreasing gate length is obvious.

Comparing the characteristics of Figures 7.41 and 7.42, it is also clear that the 22 nm device has a threshold voltage of -1.2 V, whilst the 50 nm device has a threshold voltage around -0.4 V. This would also be expected as a result of improper aspect ratio scaling.



(a) Output characteristics

(b) Transfer characteristics (V_{ds} is swept from 0.2 - 0.8 V)

Figure 7.42: Output and transfer characteristics of 3rd-generation 50 nm 25 μm-wide HEMTs.

Wide variations in device characteristics were also noted between devices of a given geometry from different cell locations across the sample, as evidenced by Figure 7.44. These 30 nm gate length devices were measured from locations at diagonally opposite sample locations. The first indicates a peak saturation drain current of over 800 mA/mm at zero gate bias and a drain voltage of 1.4 V and a peak transconductance of approximately 1 S/mm at a gate bias of -0.2 V and a drain bias of 800 mV.

The second device, designed and processed identically, indicates a zero-gate bias I_{dss} of approximately 600 mA/mm and an equivalent peak transconductance of close to 1.4 S/mm at a gate bias of 0.1 V.

The device with higher drain currents therefore exhibits lower transconductance and *vice versa*.

TLM data were also extracted from the sample. A capped contact resistance of 0.065 Ω.mm and a sheet resistivity of 119.8 Ω/sq were extracted: significantly lower figures than extracted from previous samples.

The s-parameters of the devices were also measured from 0-67 GHz.

Known physical circuit parameters were used to fit the model, leaving fewer variables to be optimised. In particular the access resistances have a strong role in both the model fit accuracy and the projected performance of the de-embedded circuit. As a

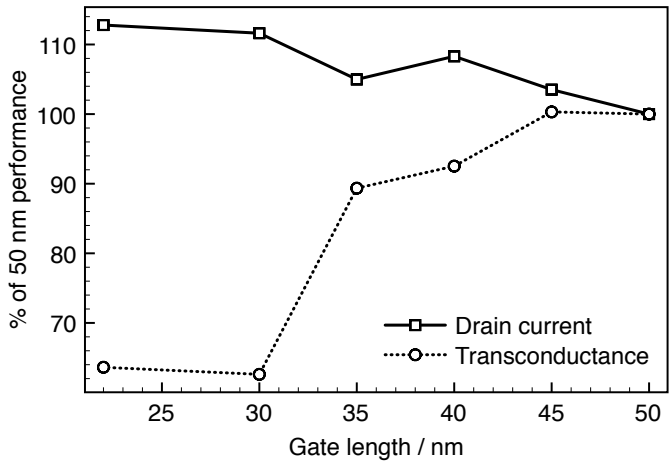


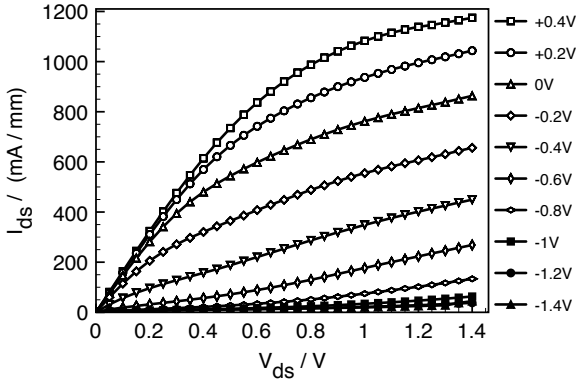
Figure 7.43: Relative drain current and transconductance as a function of gate length for C216 devices. It is probable that the 30 nm device has a spuriously low transconductance figure.

consequence, the resistivity of the material and the gate-ohmic separation were used to calculate rough values for the access resistances. These rough values were then used to fit the model to a set of measurements taken at zero applied bias, since the resistances should be bias-invariant. The extracted resistances from the best fit of the zero-bias model were then used to calibrate the access resistances of the peak-transconductance model. It is important to realise that, unlike the resistances, other model parameters are bias-variant, and as such, describe transistor action.

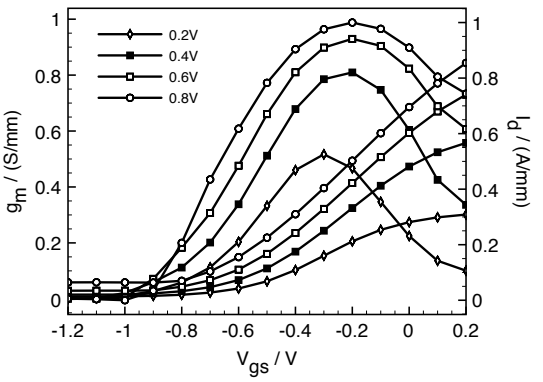
Given the sheet resistances detailed in Table 7.2, the expected sum of source and drain resistances for a standard 1.9 μm separation, neglecting the short recessed region where resistance will be higher, was on the order of 10 Ω. This was confirmed by a model fit to the unbiased s-parameters. The value of the gate resistance was also extracted from physical measurements, as from Figure 7.19, and a d.c. resistance of 45 Ω can be expected for the 300 nm gate head used in these devices. At r.f., a value of roughly one third of this is expected [81] due to its distributed nature. The source and drain changes are sufficiently minimal as to render their influence negligible [81].

The parametric fit to the measurements from the 22 nm device of Figure 7.41 is shown in Figure 7.45.

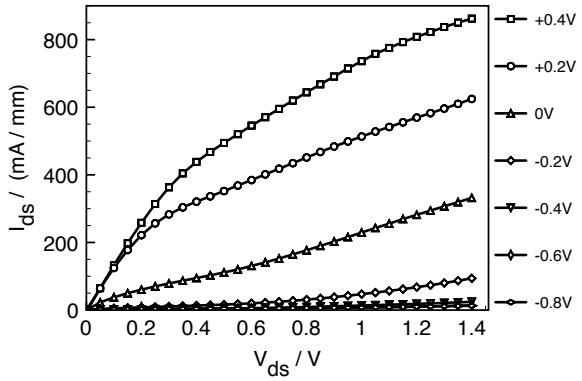
De-embedded measurements from the 22 nm device were extrapolated to high frequency to extrapolate f_t and f_{max}. As indicated in Figure 7.46, the extracted cutoff frequency



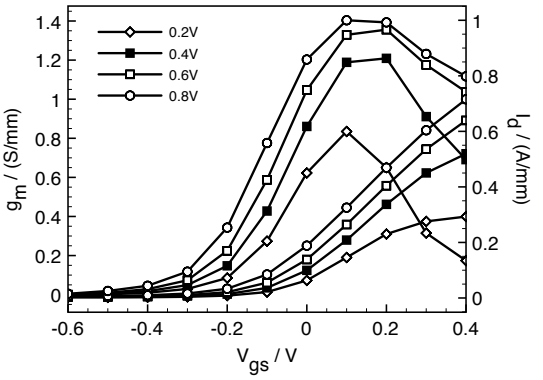
(a) Output characteristics - Device 1



(b) Transfer characteristics (V_{ds} is swept from 0.2 - 0.8 V) - Device 1



(c) Output characteristics - Device 2



(d) Transfer characteristics (V_{ds} is swept from 0.2 - 0.8 V) - Device 2

Figure 7.44: Comparison of the output and transfer characteristics of 3rd-generation 30 nm 50 μm-wide HEMTs. Device 2 features lower drain currents but improved transconductance.

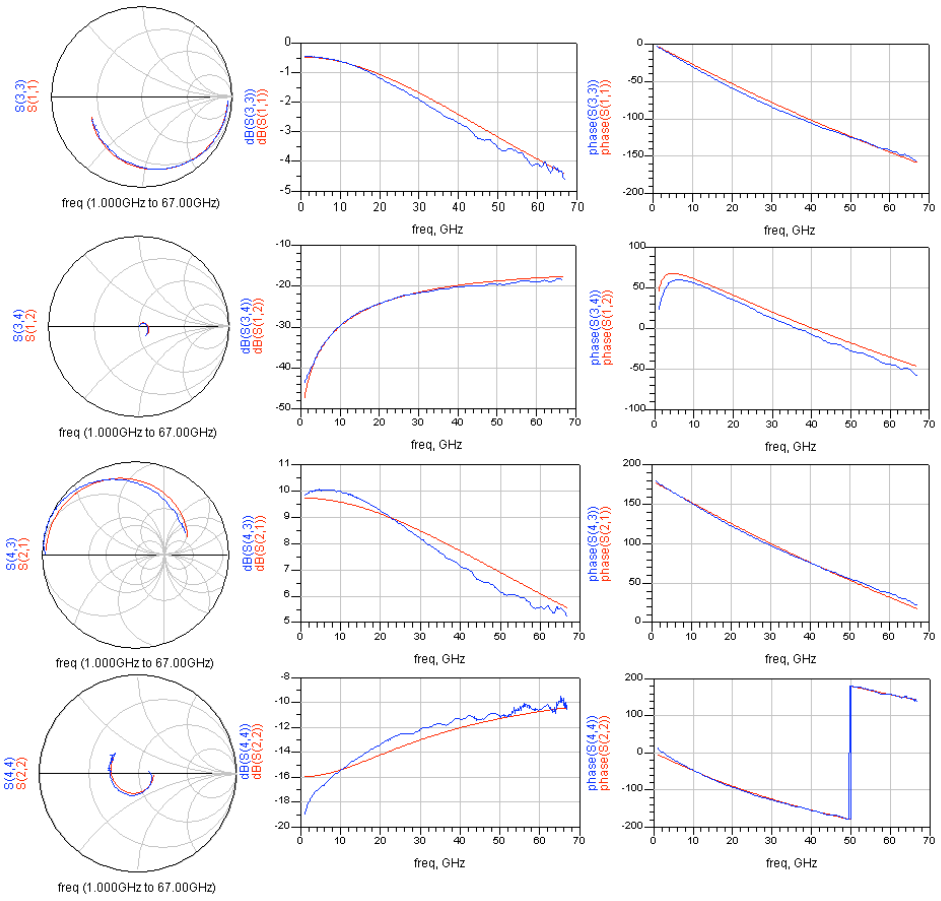
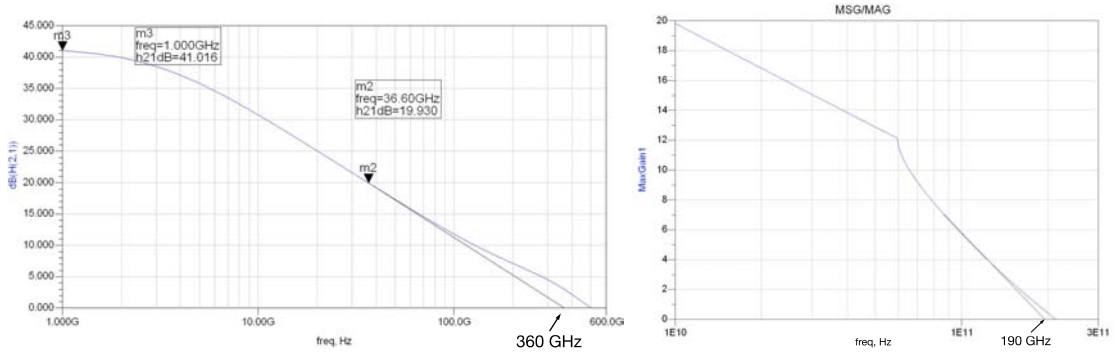


Figure 7.45: S-parameter matching of the equivalent circuit model to the 22 nm device of Figure 7.41.

was approximately 360 GHz, the maximum frequency of oscillation around 190 GHz.



(a) Extrapolation of f_t from s21 (b) Extrapolation of f_{max} from MAG/MSG

Figure 7.46: Determination of cutoff frequency and maximum frequency of oscillation from de-embedded equivalent circuit model.

The component values extracted from the equivalent circuit model are summarised in Table 7.7.

Parameter	22 nm	50 nm	Parameter	22 nm	50 nm
C_{gs}	18.5 fF	28.5 fF	R_i	0.01 Ω	1 Ω
C_{gd}	7.4 fF	8.2 fF	C_{gsp}	1 fF	1fF
C_{ds}	27 fF	18 fF	C_{gdp}	1 fF	1 fF
R_{ds}	38.5 Ω	258 Ω	C_{dsp}	1 fF	0.01 fF
g_m	0.055 S	0.09968 S	L_g	1 pH	1 pH
R_d	5.3 Ω	3.2 Ω	L_d	3 pH	3 pH
R_s	5.4 Ω	3 Ω	L_s	6 pH	6 pH
R_g	20 Ω	11 Ω			

Table 7.7: 22 nm and 50 nm equivalent circuit parameters.

By similar processes, the extraction was also done for the 50 nm device of Figure 7.42. The extracted f_t was only around 190 GHz, whilst f_{max} was around 109 GHz. The large variation is explicable when the small-signal circuit components of Figure 7.7 are compared.

Though the 50 nm device exhibits an approximately 60 % larger normalised d.c. transconductance than the 22 nm device, the capacitances extracted from the model are also much larger - C_{gs} is 10 fF greater in the 50 nm case. The 50 nm device is 25 μm wide, whilst the 22 nm device is 50 μm wide. As a result, the extracted normalised total $C_{gs} + C_{gd}$ of the 50 nm device are around more than double those of the 50 nm device. This is not

unexpected, since the gate footprint of the 50 nm device is clearly more than double that of the 22 nm device.

Considering Equation 3.60, the reason for the reduced cutoff frequency is clear: the 50 nm gate capacitances are far greater per unit area than the increase in the transconductance, degrading the cutoff frequency. It is also worth noting that the source-drain resistance, related to the output conductance, is much greater in the 50 nm case. This is also visible from the plots of Figures 7.41 and 7.42.

7.10.1 Discussion

The fabricated devices are promising in that a large percentage of the short gate length devices behaved as expected, with negligible deviations in the output characteristics and respectable d.c. figures of merit. This implies that the new gate process and process flows are not inherently damaging to the material, and the problems of earlier process runs are largely remedied. Performance variations between cells, however, were very large across a single sample, implying a processing variation.

RF measurements were additionally less promising, being much lower than previously-fabricated 50 nm pseudomorphic devices on indium phosphide.

It is particularly noteworthy that the devices have extremely high source and drain resistances, though this is in keeping with the measured sheet resistances. To examine the effect on the figures of merit, both resistances were halved in the de-embedded equivalent circuit model. Since the resistances should have only a second-order effect on the extrinsic cutoff frequency, via the effect of source resistance on transconductance, the maximum oscillation frequency would be expected to be impacted to a greater extent. Keeping all other parameters unchanged, f_t increased to 450 GHz and f_{max} to over 350 GHz.

Several additional characteristics were noteworthy between the functional devices:

1. Threshold voltage generally increases in magnitude with gate length. Shorter devices are harder to pinch off.
2. Shorter devices have decreasing transconductance in comparison to 50 nm devices.
3. Wider devices have larger output conductance than narrow devices.

4. Devices with high transconductance exhibit lower saturation drain currents and vice versa.

5. There is a wide variation of drain current, transconductance and threshold voltage across the sample.

6. There is minimal change in normalised saturation drain current within a device cell.

The problems appear two-fold:

- The material is not ideal. Firstly, it is improperly scaled for the short gate lengths considered here, as expected when the material was designed. It is therefore unsurprising that short gate length devices based on these wafers should exhibit substandard performance. Secondly, however, the access resistances are extremely large in both sets of devices: as discussed, this has a significant performance implication. Although the 50 nm devices exhibited lower performance than the 22 nm devices, this is due to the difference in capacitances alone; both devices produced disappointing performance on this material, likely due to the large sheet resistances of the new wafers when compared to earlier wafers such as A1941.

- The remaining points listed above are largely explained by variations in alignment of the gate foot to the recess, easily the most demanding process step in terms of lithographic alignment requirements. If the alignment of one to the other varies significantly, field distribution throughout the channel can be significantly altered, with large implications for electron dynamics. This may therefore explain the variations in drain current and transconductance. It also explains the lack of drain current variation within a cell.

The material must therefore be developed. This material has resulted in acceptable non-annealed contacts from a very thin cap, but high access resistances, a problem which may lie partly in the axial transport properties of the delta-doped cap. Channel transport was excellent, though improvements in its resistivity would improve matters. In addition, the gate-channel separation must also be optimised for short gate length devices. From this perspective, since 22 nm devices with cutoff frequencies well in excess of 300 GHz were realised from sub-standard unscaled material, the prospects for future device processing from these process modules are promising.

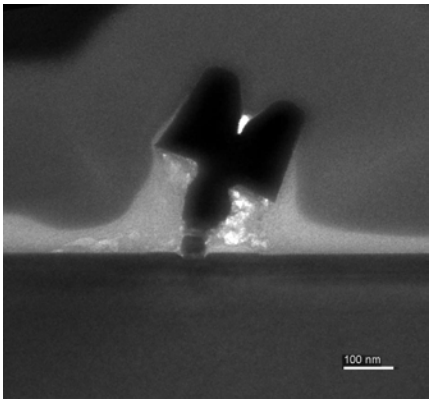
The lithographic alignment accuracy of the system is also a major obstacle. Although the VB6 has vastly superior accuracy to previous EBL systems, the basic system is inadequate for high-uniformity processing on this scale.

In order to verify the layout of the devices, with particular emphasis on investigation of the alignment of the gate to the recess, it was decided to investigate the functional devices using Transmission Electron Microscopy (TEM). The samples were prepared by Focussed Ion Beam milling (FIB) using an FEI Nova 200 Dualbeam system and lifted out *in situ* using an Omniprobe micromanipulator. They were then imaged using an FEI Tecnai T20 microscope.

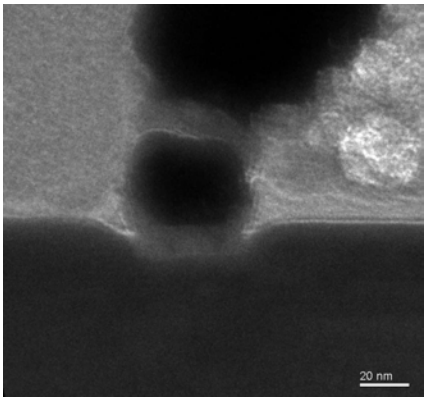
This preparation process is more appropriate to the imaging of these devices and any future structures, for several reasons. Firstly, the dimensions of interest are at the extreme capabilities of the highest-resolution SEM facilities available in the university, since surface layers of oxide and interfaces between materials are of interest. The length scales of interest are in fact therefore smaller than the shortest lithographically- or epitaxially-defined structures, but are sub-nanometre. Secondly, as previously discussed in Section 7.7, the silicon nitride film has inherent stress which makes imaging after conventional sample preparation by cleaving very difficult. In addition to charging effects, traditional SEM approaches are unsuitable. The FIB/TEM methods, however, yield a far flatter and undamaged milled surface to be imaged, and are intrinsically capable of much greater resolution.

The samples imaged, shown in Figure 7.47, had the silicon nitride etched off following device measurement, however, some of the film remained under the gate head. Several conclusions can be drawn from these images. Firstly, the gate is not placed in the centre of the defined recess, as is clear from Figure 7.47(b) and as it ideally would be. It is therefore clear that misalignment may play a major role in device variations. In addition, as can be seen from the high magnification image of Figure 7.47(c), the recess has extended below the silicon nitride, though the cap edge itself is very nearly touching the gate.

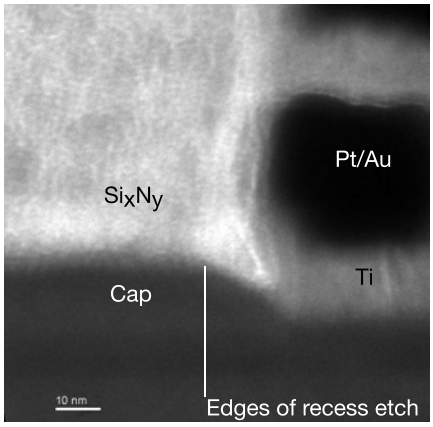
As a consequence, it would appear that the recess etch has yielded a smaller trench than intended, and the pre-evaporation succinic acid etch has resulted in undercut of the dielectric and additional removal of the material. Given the asymmetry observed in the previous paragraph, the second etch step would not appear to be the only active recess process. The recess process, therefore, though functional, needs additional calibration to ensure uniformity.



(a) Overview of gate region



(b) Overview of gate recess. The gate is not placed symmetrically in the recess.



(c) High magnification view of recess edge. The recess etch has extended under the silicon nitride.

Figure 7.47: TEM images of 50 nm devices, showing gate region after silicon nitride etching.

The third-generation devices investigated in this TEM study had the nitride removed after measurement using a blanket-etch SF_6 process. The devices were then re-measured to investigate the effects of silicon nitride removal on the device performance, since the reduction in dielectric constant should reduce the total gate capacitances. Previous tests had shown that the silicon nitride can be removed after device completion by either wet or dry etch processes using HF or SF_6. After removal of the gate dielectric support, the gates are extremely delicate, since they become self-supporting; hence wet-etched devices suffered poor gate yield after etching as a consequence of the forces exerted by immersion in liquid. Coupled with the transport damage effects observed in the use of HF in InAlAs processing (Section 7.9), dry etch processes appeared intrinsically more suitable and yielded more mechanically stable gate structures.

Investigation of various etch times concluded that over-etching was required to remove the dielectric from under the gate, and a time of 4.5 minutes was found to remove most of the silicon nitride film, with slight residue remaining under the gate. As a result, a 5 minute etch time was used to strip the film prior to measurement. Unfortunately, this etch time proved damaging to the exposed surfaces and resulted in severe degradation of the device performance, though the bulk of the silicon nitride was removed.

RF measurements of the devices were used to extract the equivalent circuit parameters for comparison with those of the same devices prior to etching. The extracted parameters are shown for the same 22 nm device measured previously in comparison before and after etching in Table 7.8. These yielded a de-embedded cutoff frequency of only 137GHz, highlighting the extreme effect of the etch damage on a device which previously achieved an f_t of 360 GHz.

Parameter	Unetched	Etched	Parameter	Unetched	Etched
C_{gs}	18.5 fF	10.4 fF	R_i	0.01 Ω	1 Ω
C_{gd}	7.4 fF	7.0 fF	C_{gsp}	1 fF	1fF
C_{ds}	27 fF	19 fF	C_{gdp}	1 fF	1 fF
R_{ds}	38.5 Ω	105 Ω	C_{dsp}	1 fF	0.01 fF
g_m	0.09968 S	0.05368 S	L_g	1 pH	1 pH
R_d	5.3 Ω	58.8 Ω	L_d	3 pH	3 pH
R_s	5.4 Ω	22.3 Ω	L_s	6 pH	6 pH
R_g	20 Ω	24.9 Ω			

Table 7.8: Effect of 5m silicon nitride etch on 22 nm equivalent circuit parameters.

As is evident from the table, there are two key differences in device equivalent circuit components. Firstly, the total source - drain resistance has increased by a factor of eight,

while the output resistance is similarly increased. Secondly, the capacitances associated with the gate (modelled here by C_{gs}, C_{gd} and C_{ds} only for simplicity of modelling) have decreased by 30 %. The etch process therefore appears to have significantly reduced the capacitances due to the dielectric film, though the over-etching has markedly damaged the underlying semiconductor.

The etch process clearly needs further optimisation, given the large extent of the damage and the residual silicon nitride under the gate in Figure 7.47(a). To consider the effects of the capacitance reduction on device performance, assuming damage issues with the etch can be resolved, the post-etch capacitances were used in place of those of the initial 22 nm device parameters and the circuit remodelled. Without changing other parameters, the cutoff frequency increased to 480 GHz and f_{max} to 360 GHz from their initial figures of 360 GHz and 190 GHz, respectively.

Including the previously-mentioned halving of resistance which might result from epitaxial optimisation, the circuit was again simulated. The resultant f_t was over 600 GHz, and f_{max} over 450 GHz.

Removal of the silicon nitride therefore appears to be a worthwhile avenue of exploration in the enhancement of device performance. The etch process will, however, need considerable further optimisation. Furthermore, the equivalent circuit results underscore the need for optimisation of all parasitic elements

7.11 Summary

This chapter has described the development of a gate module suitable for the high-yield fabrication of sub-25 nm gates. The process was designed to eliminate as completely as possible the problems associated with traditional single-step gate processes: namely, pattern definition issues related to electron scattering in thick resist, metal evaporation and mechanical stability issues related to increasing resist aspect ratio and issues of increasing gate length extension into a self-aligned recess trench. The solution developed involved the high-resolution electron beam lithography and reactive ion etching of a thin silicon nitride film, deposited by room temperature inductively-coupled plasma chemical vapour deposition.

The key changes to the process were the decoupling of the gate foot and gate bulk lithography, and the definition of the gate recess in a separate lithographic step. The

three key process parameters are realised in three distinct process stages, allowing each to be tailored individually. The result is completed devices where the entire semiconductor surface between the gate and ohmic contact is encapsulated in silicon nitride and there is great liberty in defining the device geometry to tailor performance.

In incorporating the new processes into a complete, re-ordered process flow, several process steps required to be created or optimised. In particular, surface treatments and de-oxidation steps were shown to be of particular importance in defining the resultant device output characteristics, with aggressive acid cleans proving particularly degrading to device characteristics.

Both double and single delta-doped material was designed and grown metamorphically on gallum arsenide using MBE. The material featured a new cap doping strategy to allow non-annealed contact formation on a very thin cap to improve prospects for gate scaling. Contact resistances were very slightly higher, whilst sheet resistances were considerably higher than previous bulk-doped structures as a consequence of the new strategies. The material grown initially was additionally not sufficiently scaled for use as an ideal layer structure for ultra short gate length devices; rather, it was intended as an evolution of previous material for use as a test bed for the new cap structure on which future scaled layers might be based.

After two initial device runs in which processing problems were identified, functional devices were realised across a spread of gate lengths. These devices exhibited excellent d.c. performance, with drain currents up to 1.2 A/mm and transconductance over 1 S/mm for the 22 nm devices, whilst 50 nm devices yielded transconductances of up to 1.6 S/mm with drain currents above 1 A/mm. A general trend of decreasing transconductance with gate length was observed, indicating that, indeed, the material was insufficiently scaled for the gate lengths realised using the new processes. In addition, the resistivity of the material proved to be too high, resulting in high access resistances, to the detriment of device r.f. performance, which yielded a cutoff frequency of 360 GHz and a maximum frequency of oscillation of 190 GHz for the 22 nm devices.

The devices also showed unintended variations in characteristics across the sample, suspected to be related to slight differences in alignment accuracy between device cells.

The device results thus confirmed the suitability of the developed process flow in the fabrication of devices with gate lengths of less than 25 nm. Both the material design and alignment process, however, need further examination if devices with leading performance

are to be realised.

8. Device Development

8.1 Introduction

Sub-25 nm devices were successfully realised using the two-step processes outlined in Chapter 7. Key issues with the material design and alignment accuracy were noted to be the limiting factors to the performance of the final devices. As a consequence, these were key areas to be explored in the development of further devices.

The devices fabricated in the previous chapter featured a fairly conservative approach to the two-step gate process, with relatively thick resists and silicon nitride used for the realisation of 22 nm gates. As a consequence, having verified the feasibility of the process for short gate length device fabrication, the two-step strategy was expected to be scaleable to produce shorter gate lengths by employing more aggressive fabrication strategies in the definition of the gate foot.

This chapter details the progress made in advancing the fabrication techniques introduced in Chapter 7 to produce truly aggressive device geometries.

8.2 Alignment techniques

The 22 nm devices fabricated suffered from large process variability as a consequence of variation of the position of the gate foot with respect to the recess as discussed in Section 7.10. Standard cell alignment techniques used in electron beam lithography had proved to markedly improve alignment accuracy over the simple case of using global markers only. It was clear, however, that a further increase in accuracy would be required to ensure process stability.

Standard mark locate strategies have serious limitations, since they extract information

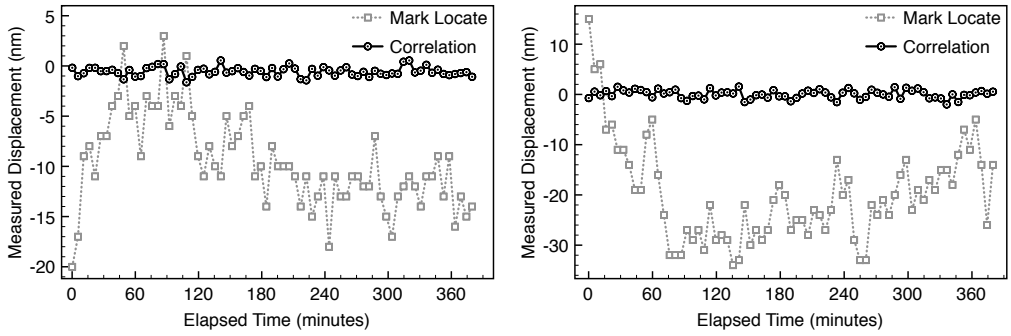
from only the marker edges, resulting in errors in position measurement due to rotation or roughness. Correlation-based alignment using image capture techniques provides a mechanism for extracting much more information, allowing the centre of the marker to be located by correlating the captured image to an idealised image. The marker images can be captured by rastering the electron beam across markers and measuring the secondary electrons, as in an SEM.

Research into the development of correlation-based alignment techniques was carried out in the department by K. Docherty in parallel with the work reported in this thesis. Routines for alignment were created for the departmental Vistec VB6, allowing the use of correlation-based alignment in generic lithography jobs. Various marker types were also investigated for their effectiveness during correlation. In particular, markers with high pattern density were used to provide the maximum number of reference points. Aperiodic markers of varying density proved to be optimal, and Penrose tilings, which form entirely aperiodic patterns were shown to be a particularly effective pattern choice [336]. By correlation of a Penrose tile of approximately 8 μm in extent fabricated in Ti/Au by lift-off, correlation-based methods proved to yield alignment accuracy around ten times better than conventional alignment methods, as shown in Figure 8.1.

The potential combined alignment accuracy in x and y was found to be less than 1 nm, compared with 11 nm for the conventional mark locate. It is noteworthy that this latter figure corresponds well with the 15 nm alignment accuracy extracted for the mark locate cell alignment in Chapter 7. It therefore seems probable that the correlation-based method might yield a solution to the variability problems found in the short gate length HEMTs.

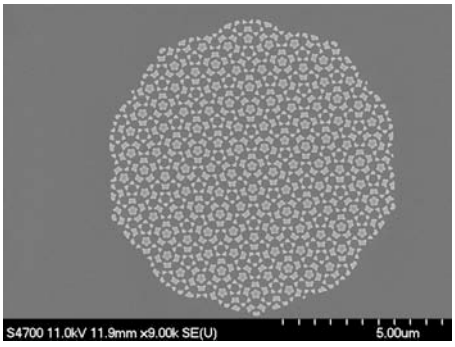
As a consequence of the high pattern density and relatively small feature size of around 100 nm used in the Penrose patterns, the markers required to be fabricated using thin resist, hence thinned metals. This presents no problems for the two-step process developed, since the marker level must already be defined independently prior to recess etching. As a result, the patterns were dose-tested in PMMA on bulk GaAs. A completed marker is shown in Figure 8.1(c).

Given their superior alignment performance, the Penrose markers were substituted for the square cell markers in subsequent device substrates.



(a) x-axis.

(b) y-axis.



(c) SEM micrograph of completed Penrose tile on GaAs.

Figure 8.1: Comparison of mark locate and correlation-based alignment methods for accuracy. After [336]. Also shows a lifted-off Penrose marker.

8.3 10 nm gate process development

The two-step method described in Section 7.4 relies on the definition of the gate foot by electron beam lithography and anisotropic SF_6/N_2 reactive ion etching (RIE) of a thin layer of silicon nitride.

A single layer of ZEP520A electron beam resist is used for pattern transfer into the silicon nitride, with the bulk of the gate structure aligned separately. Although the two-step method [337] requires a second lithographic exposure, it allows considerable flexibility in defining the foot layer, removing many of the processing constraints associated with a single-step process. The process was expected to be capable of further development to yield features much smaller than those realised by the conservative initial process flow.

This section aims to details progress in further development of the two-step process by

the modification of only the foot definition step. The upper gate step can then remain unaffected, fabricated on the surface resulting from the foot definition as previously. As previously, all foot definition occurs after the fabrication of the recess trench.

As a consequence of the flexibility afforded by the two-step process, two separate methods were investigated for the definition of very short features. The first relies on the reduction of forward scattering inherent to the use of thin resists and optimisation of the exposure, development and pattern transfer processes. The second focusses on the reduction of the dimensions of an initially-defined feature using plasma processing. Though the two approaches are very different in methodology, both have the same goal of a resultant metallised 10 nm gate foot encapsulated in silicon nitride, suitable for subsequent definition of the upper gate head.

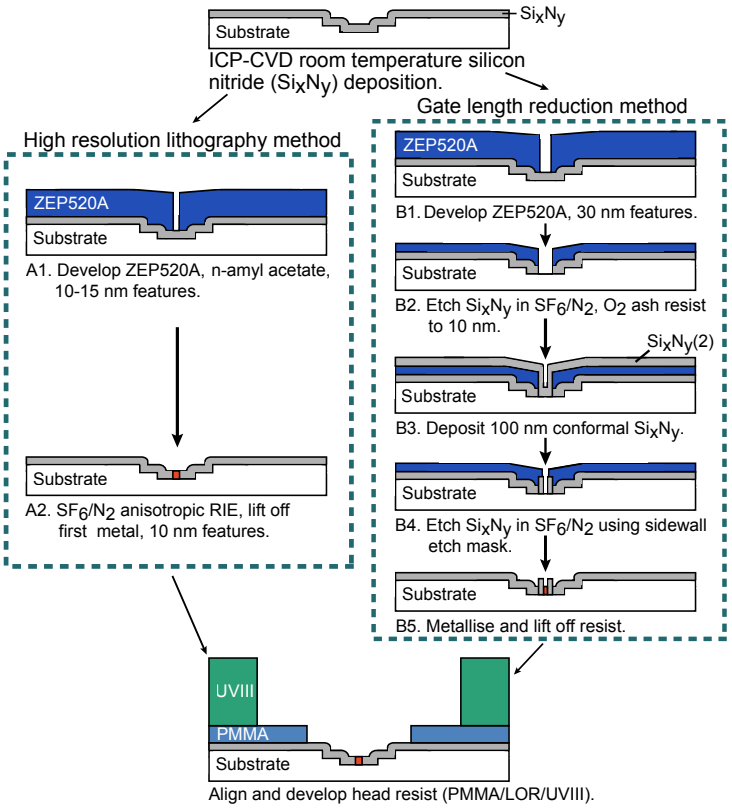


Figure 8.2: Overview of high resolution and gate length reduction methods.

An overview of the two processes is shown in Figure 8.2.

8.3.1 High resolution technique

Thin ZEP520A has previously exhibited isolated lines on the order of 10 nm [154] on silicon, but few references exist for its attainable resolution on III-V substrates. It has previously been reported that the n-alkyl-acetate family of developers can yield improved resolution and contrast in ZEP520A [338], whilst developer molecular weight has an effect on residue in developed areas [153].

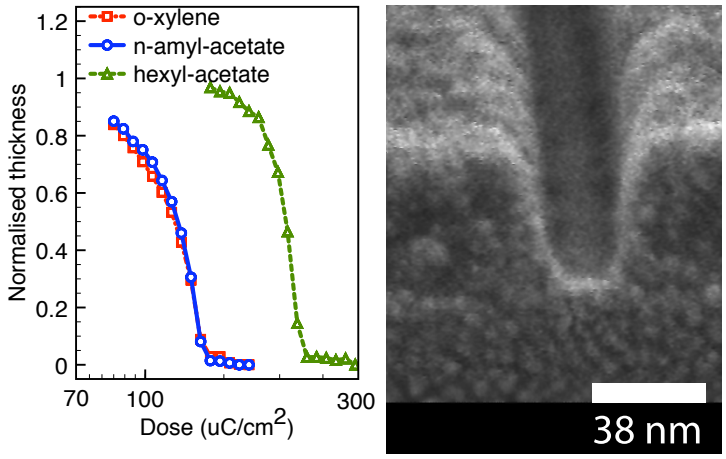
To investigate the performance of thin ZEP520A when developed using various different developers, the contrast curves were extracted for each developer solution for the same film thickness. 50 nm-thick ZEP520A was spun on several samples of bulk GaAs and a grating of 2x2 μm squares was dose tested using a spot size of approximately 4 nm, the smallest spot possible using the VB6. The samples were then developed for 30 s in o-xylene, amyl- and hexyl-acetate respectively and rinsed in isopropyl alcohol. O-xylene was the developer used previously for 22 nm gates, for purposes of comparison with the alkyl acetates. The relative residual resist thickness was then characterised by AFM as shown in Figure 8.3(c) for each of the relevant exposure doses, allowing contrast curves to be obtained for the three developers.

The contrast curves extracted are shown in Figure 8.3(a). Both amyl- and hexyl-acetate produced markedly improved contrast over o-xylene, with hexyl-acetate providing relatively best contrast overall.

Single-pixel lines were then exposed using the same process and examined in cross-section by SEM. Using o-xylene, it proved impossible to realise well-cleared features smaller than around 20 nm; a situation similar to the development of a 100 nm-thick resist film as discovered for the 22 nm process. Hexyl-acetate yielded best contrast, and whilst 10-15 nm features were achieved, residue was evident across the dose range. This finding is in keeping with the work of Yamaguchi, et al. [153].

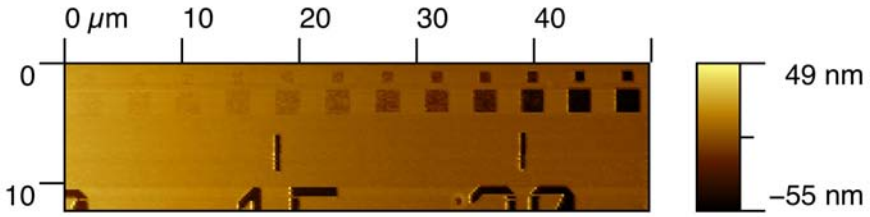
Amyl-acetate, however, produced well-cleared, high-contrast features on the order of 10-15 nm, as shown in Figure 8.3(b). This resist was then used as an etch mask for SF_6/N_2 RIE [256] of a 20 nm-thick ICP silicon nitride (Si_xN_y) film deposited at room temperature for various etch times. The determination of the resultant dose test linewidths was then required for each etch time.

As previously discussed, the imaging of small structures in silicon nitride is extremely difficult as a consequence of sample charging, film stress and the difficulty of operating



(a) Contrast curves resulting from the development of 50 nm ZEP520A in o-xylene, amyl- and hexyl-acetate.

(b) 10-15 nm features in ZEP520A after amyl-acetate development.



(c) Contrast curve structures measured by AFM after development.

Figure 8.3: Contrast curves measured for the n-alkyl acetates.

at the extreme resolution of the microscope. As a consequence, it was decided to image the samples by TEM as previously used for the analysis of later devices. As previously, the samples were prepared for TEM examination by Focussed Ion Beam (FIB) milling. The ion/electron beams were used to deposit a protective layer of platinum over the area of interest to prevent sample damage during milling.

Figure 8.4 shows the dependence of the resultant feature size on dose and etch time as measured by averaging several sites of the same dose on each sample. A 90 s etch time produced features as small as 4-5 nm for low doses, but these would be highly dependent on precise resist geometry and etch termination, and hence unsuitable for realistic device applications. Moving to slightly higher doses, however, produced repeatable, well etched 10 nm trenches, as shown in Figure 8.5(a).

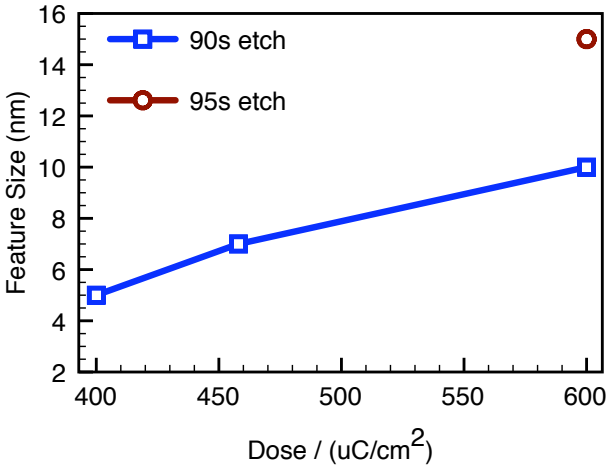


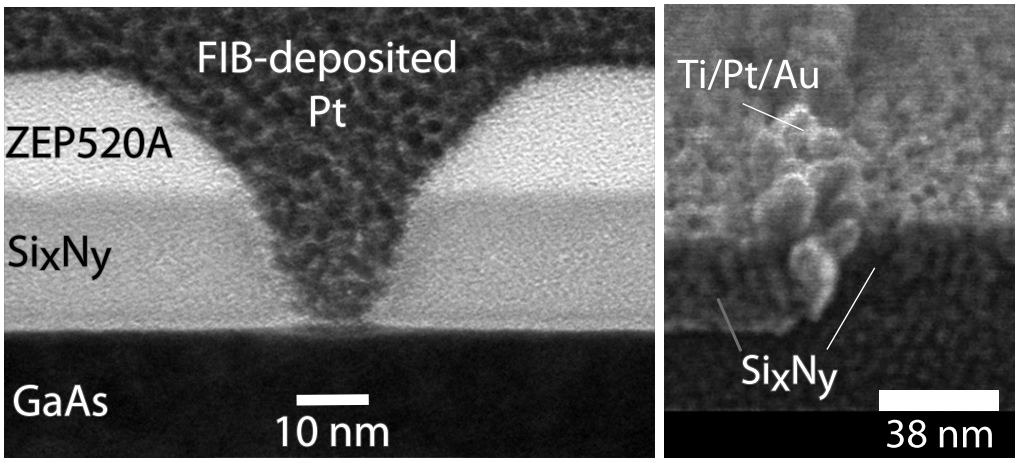
Figure 8.4: Dependence of etch trench dimension on exposure dose and etch time.

The etched dimensions are hence extremely sensitive to over-etching, with an additional 5 nm lateral etch resulting from a 5 s extension of etch time; therefore control of this process is crucial. It is useful to note the geometry of the etched trenches as shown in Figure 8.5(a)

A thin gate metallisation of Ti/Pt/Au was then evaporated into the etched trench and lifted off using the residual ZEP520A as a mask, leaving a relatively planar surface for the upper gate lithography described previously. The upper level was then aligned to the etched trench and lifted off, realising complete 10 nm gates as shown in Figure 8.6.

As a consequence of the low contrast between the gate foot metal and the silicon nitride on such small length scales in the SEM, dashed lines have been added as a guide to the eye in Figure 8.6(a). The dark triangular region on top of the silicon nitride represents the area above the etched trench, where excess metal from the gate foot liftoff has formed.

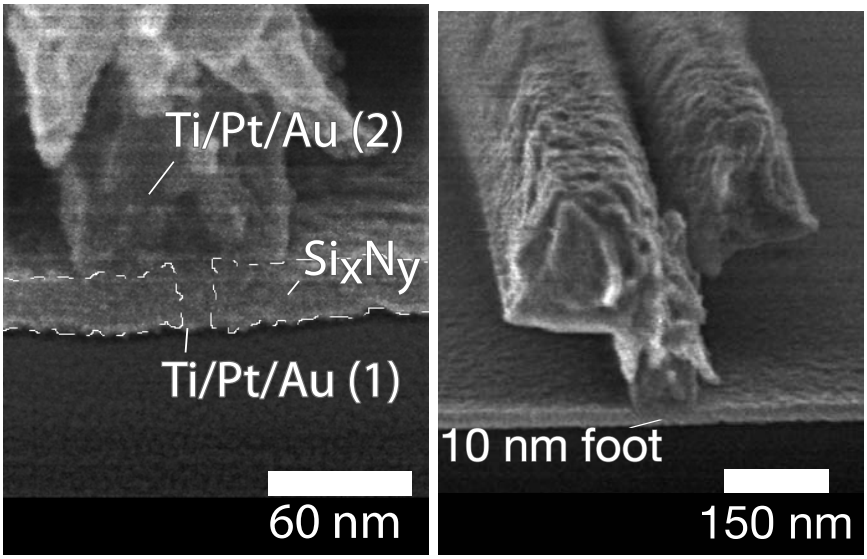
The high resolution method is hence a feasible method for the fabrication of gates as small as 10 nm. The method requires precise control over the conditions of the resist, development and etch, but well-controlled processes have resulted in the transfer of features as small as 4-5 nm in some cases. The process appears to be a viable method for 10 nm T-gate fabrication when higher doses are used, where resultant etch profiles showed high uniformity.



(a) TEM image of 10 nm etch trench resulting from SF_6/N_2 RIE of silicon nitride using ZEP520A etch mask of Fig. 2. Platinum is deposited by ion beam deposition to protect the sample during milling.

(b) Metallised 10 nm gate foot fabricated by liftoff using remaining ZEP520A as a mask.

Figure 8.5: 10 nm etch trench in silicon nitride and lifted-off gate foot profile using the same etch trench.



(a) Completed 10 nm T-gate fabricated by the high-resolution method - gate foot is shown below dark triangular region. Dashed guidelines have been added for emphasis of contrast.

(b) Low magnification image of complete gate of 8.6(a), where the foot is virtually indistinguishable due to its dimensions.

Figure 8.6: SEM images of completed 10 nm gates.

8.3.2 Length reduction technique

The second method for short gate length fabrication is the reduction of features produced by the conventional process by the use of plasma processing. Whilst modern electron beam lithography tools are capable of high resolution, it may be desirable to minimise the processing constraints associated with the fabrication of extremely small features. Consequently, a gate length reduction technique based around larger lithographic features was developed.

Plasma-based dielectric deposition and etch processes are frequently used in the silicon industry to reduce the feature sizes produced by optical lithography. Sidewall spacer processes [255, 339, 340] for self-aligned ohmic contact formation or etch-back processes [341] for gate patterning below the lithographic limit are good examples. These methods generally rely upon the potential for conformal deposition afforded by chemical vapour deposition methods, together with the vertical etch processes possible using RIE techniques.

The method chosen was similarly based around the same principles, using the damage-free silicon nitride deposition and etch processes detailed in Chapter 7. A trench is defined in ZEP520A and transferred to silicon nitride by the usual SF_6/N_2 RIE process. The residual resist then remains in place after etching. A layer of silicon nitride is then deposited on top of the resist / silicon nitride bilayer, with conformance resulting in deposition of silicon nitride into the defined trench and onto the trench sidewalls. Vertical silicon nitride walls are therefore formed in the trench foot, with a short gap between them in the foot.

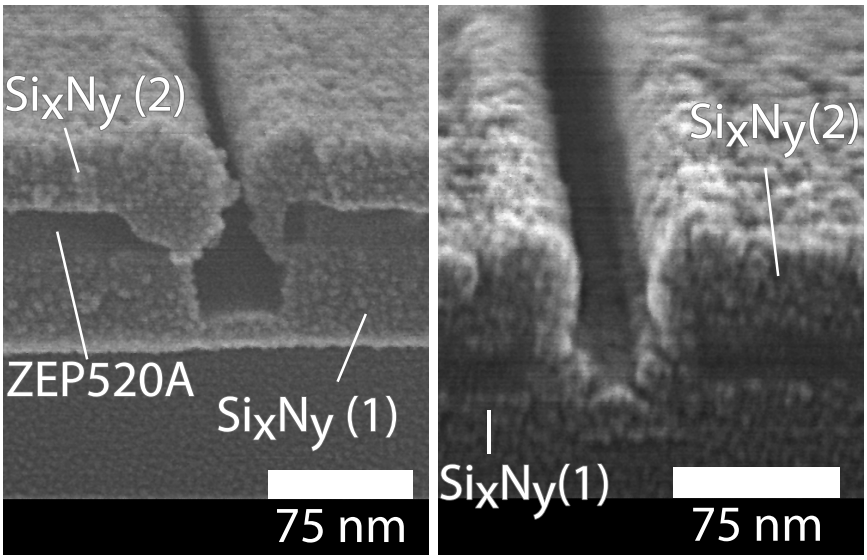
The second deposition should therefore be relatively thin in the foot of the trench, with high aspect ratio vertical sidewalls. The proposed process makes use of the short gap and thick sidewalls. By exposing the complete structure to a second anisotropic RIE step, therefore, the bulk silicon nitride should be etched reasonably rapidly, whilst the vertical sidewalls, exposed to a perfectly vertical etch process, should be resistant to the etch.

As a consequence, an optimised etch time and dielectric deposition should yield a reduced feature size in the trench foot. The gate foot can then be metallised and lifted off using the residual ZEP520A after completion of the RIE process. The full process is outlined in Figure 8.2.

A 30 nm trench was fabricated in 20 nm-thick silicon nitride by the exposure of 100 nm

ZEP520A, development in o-xylene and SF_6/N_2 RIE as per the data of Figure 7.13.

In initial processing tests, relatively thin films of 15-35 nm were deposited on top of a relatively thicker underlying silicon nitride/ZEP520A stack of around 100 nm. Cross-sectional SEM was used to assess the conformance and etching of the second layer of silicon nitride. It is clear that the thickness of the deposited silicon nitride and the aspect ratio of the underlying trench are extremely important. If the aspect ratio of the silicon nitride and remaining resist is too high, deposition of the second layer of silicon nitride becomes discontinuous regardless of the thickness of the layer, as is clear in Figure 8.7(a).



(a) Deposition of 25 nm silicon nitride onto high-aspect ratio etch trench, resulting in discontinuous sidewalls.

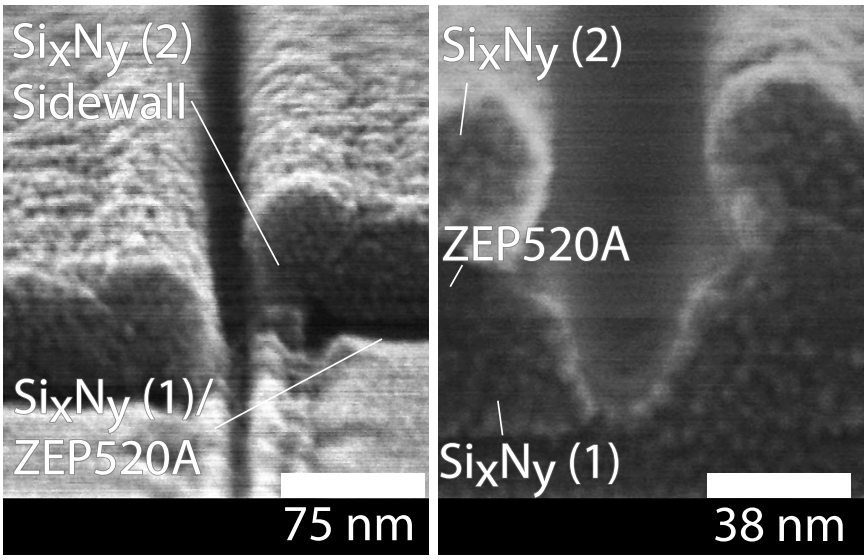
(b) Non-vertical sidewalls on ashed bilayer due to insufficiently thin film.

Figure 8.7: Sidewall formation issues in the length reduction process.

To counter this problem, the remaining resist was thinned to approximately 10 nm using a low-damage O_2 RIE process [342], reducing the aspect ratio after ashing to around 2:1. Various thicknesses of ICP silicon nitride were then deposited on top of the ZEP520A/Si_xN_y bilayer. The film thickness remains extremely important, since excessively thin films result in the formation of non-vertical sidewalls which would be unsuitable for etching of the foot, as is clear in Figure 8.7(b).

By systematic variation, the optimal thickness was found to be 100 nm, which resulted

in continuous vertical sidewalls and complete filling of the etch trench (Figure 8.8(a)). This conformance was repeatable across eight sites examined over two samples, yielding vertical sidewalls suitable for etching.



(a) Deposition of 100 nm silicon nitride onto ashed etch trench results thick vertical sidewalls and completely filled trench.

(b) 10 nm etch trench formed in second silicon nitride film by SF_6/N_2 RIE.

Figure 8.8: 10 nm trench formation after O_2 ashing.

The structure was then exposed to further anisotropic SF_6/N_2 RIE without additional masking for various etch times, allowing the silicon nitride in the foot of the trench to be etched using the sidewalls as the etch mask, yielding the structure of Figure 8.4.

An optimal 130 s etch time resulted in complete etching of the silicon nitride in the foot of the trench and retention of the vertical sidewalls, resulting in a 10 nm footprint, as in Figure 8.8(b). Feature size is, as for the high-resolution method, highly dependent on over-etch time, since the sidewalls erode with increasing etch time, though consistent etch profiles were achieved across the four sites examined for each time.

Crucially, however, reproducible short gate lengths on the order of 10 nm are achieved by only relaxed lithography using this method.

The completed structure can then be metallised and any additional silicon nitride lifted off using the remaining ZEP520A along with the extraneous gate metal, then incorporated

into the two-step process flow as for the high-resolution method.

8.3.3 Discussion

Two separate methods for the fabrication of 10 nm T-gates have been proposed, and the issues surrounding both techniques discussed. In the case of the high-resolution method, developer selection is important as part of optimal resist processing of minimally thin resist films, and the resultant linewidth is very sensitive to over-etching by RIE. Careful control of both development and etching process steps, however, can result in features of 10 nm or less.

The relative dielectric thicknesses, etch times and aspect ratios play a crucial role in film continuity and hence in achievable feature sizes in the length reduction method, though the processing requirements are significantly relaxed. In particular, the critical emphasis in processing shifts from lithographic control to repeatable plasma processing, through the use of precise etch times and film geometry.

Both methods can be incorporated into a two-step T-gate process and have resulted in the fabrication of 10 nm T-gates.

An intriguing further possibility is the combination of the two methods for the formation of repeatable nanometric gate lengths. In particular, since the "high resolution" method of Section 8.3.1 has shown limitations in repeatability at trench lengths of less than 10 nm, it is possible that if the geometry could be significantly well controlled, a length reduction technique based on the 10 nm structures resulting from a well-controlled development and first etch might be expected to yield sub-10 nm gate lengths with repeatability.

As a result, since methods exist for the definition of truly nanometric structures, the issues of realising well-scaled epitaxial structures for the fabrication of scaled devices are again highlighted.

8.4 Ohmic contact development

The 22 nm devices realised previously featured a scaled 22 nm gate centred in a 2 μm ohmic gap. For complete device scaling, however, there may be additional performance to be gained by the scaling of the ohmic gap to smaller dimensions. Indeed, work on full-band Monte Carlo HEMT simulation by Ayubi-Moak, Akis and Ferry has suggested the key role of source-drain separation in defining the ultimate performance

of a device [22]. In particular, simulations of HEMT structures have shown the drain current and transconductance to be linearly dependent on source-drain separation. It is noteworthy that these simulations are idealised, and do not include the contributions of additional parasitic elements arising from the gate or contacts. They may, however, be useful in their indication of potential intrinsic device performance.

As the ohmic contacts are brought closer together, two major effects occur in double delta doped material. Firstly, as described in Section 3.7.1, the parasitic source and drain resistances are reduced as the lateral spacing of the contacts is reduced. It is noteworthy that the reduction in the separation will also result in an increased capacitance between the contacts and the gate in addition to the resistance reduction.

A second effect, however, is the change in electric field distribution in the device channel as a consequence of the revised contact spacing. A given voltage applied over a reduced separation intrinsically implies an increased electric field magnitude. As a result, reducing the ohmic spacing should result in increased channel electron energy, with the various related phenomena of variable valley occupancy, scattering effects, impact ionisation and real space transfer. As a consequence, it is to be expected that electron velocity would be significantly affected by large reductions in ohmic separation; whether it is increased or decreased on average, however, is dependent on the velocity saturation and overshoot characteristics of the device. In addition, it would be expected that breakdown would occur at lower applied drain voltages, and potential for increased kink phenomena.

The short gate length devices fabricated during the course of this project should feature significant velocity overshoot or quasi-ballistic electron transport as discussed in Section 3.8. As a consequence, they should benefit from reduced ohmic contact separation, since a relative increase in electric field for a given applied drain voltage should have a minimal impact.

In devices where the overshoot velocity dominates the current, electric field gradients are crucial, since relaxation times determine velocity. Iincreased field gradients should therefore be beneficial, imparting electrons with maximum energy which cannot subsequently be sufficiently relaxed as to slow the electron. Quasi-ballistic devices require that the electron be injected with maximal initial velocity in order to maximise the transit velocity towards the ballistic limit, so high electric field gradients should also positively enhance transport in these devices.

As a result, it would be expected that these short gate length devices should benefit from

scaling of the ohmic separation. Previous devices have used self-alignment of the ohmic contacts to the gate to form short gaps. In the case of HEMTs, this has resulted in source-drain separations of 300 nm, but devices fabricated by this method in fact yielded only moderate increases in high-frequency performance as a consequence of the increased parasitic capacitance resulting from the fabrication technique of using the T-gate head as a mask [58]. In particular, the limitation is predominantly that the deposited ohmic contact must be thinner than the gate foot thickness minus the cap thickness to prevent shorting of the gate head to the ohmic pad. As a consequence, the two metal layers are in extremely close proximity by the completion of the ohmic metallisation, increasing capacitance.

CMOS MOSFETs use a self-aligned architecture to yield source-drain contact spacings on the order of 80 nm for 65 nm processes [343]. The sidewall spacer processes used in digital devices, however, are not suitable for use with T-gates and do not incorporate a gate recess.

The two-step process flow introduced in Chapter 7 involves the definition of the gate in two distinct steps following recess etching. The gate foot lithography in particular requires extreme processing care, allowing the shortest possible structures to be defined, as is clear from the definition of 10 nm structures using the high-resolution method of Section 8.3.1.

The separation of the gate head and foot lithography, however, affords great flexibility in defining all elements of the device, since the fragile gate head is defined separately from the foot. The two steps can therefore be separated by other lithography steps, so the process should accommodate the lithographic definition of ohmic contacts with a short source-drain spacing.

It is proposed that the ohmic contacts be defined following the gate foot definition, but prior to the gate head lithography, emulating a sidewall spacer process using the silicon nitride already used for gate foot definition. The key requirements for the process are as follows:

- Contacts must remain low-resistance.
- Contact spacing must be greater than the recess length.
- Contacts must be aligned with precision to the gate recess in particular.

- Resist must be sufficiently thick to allow etching of the silicon nitride film.

To achieve these objectives, a single-step resist process was attempted for the realisation of close-spaced ohmic contacts.

8.4.1 Issues around the fabrication of short ohmic gaps

The fabrication of two large pads separated by a uniform short gap along the pad width is extremely challenging using electron beam lithography. As discussed in Section 4.3.2, the proximity effect limits the definition of large areas by the exposure of resist by backscattered electrons from an adjacent patterned region.

As a consequence of the proximity effect, fabricating two large adjacent structures with a well-defined central unexposed region requires extremely high resist contrast. Although there are fabrication methods to improve resist contrast, such as control of developer or its temperature, the optimal method for developing short ohmic gaps would seem to be the removal of the need to define the two pads concurrently. In particular, realising a well-defined central gap whilst completely developing the resist across the whole pad region is extremely difficult, causing problems in forming a uniform ohmic contact.

The electron backscatter radius on GaAs is on the order of multiple microns [344, 345] at 100 kV. Lithography within this region should therefore result in all areas receiving a relatively constant exposure contribution from backscattered electrons.

As a result, it was decided to define the ohmic contacts as 2 μm-long pads at a small separation along the complete contact width, within the backscatter radius, then align a larger pad to the first level. As a consequence, the first pad should be relatively uniformly exposed, with the only exposure issues occurring at the ends of the pads. The exposure requirements are then significantly relaxed for the second pad level.

The ability to define an effective short ohmic contact is closely related to the contact resistance, and is defined by the transfer length of the contact. Transfer length is defined as the distance from the edge of the contact over which $\frac{1}{e}$ of the total current is transferred from the semiconductor to the metal [181], since current does not flow uniformly through ohmic contacts to semiconductors [182]. Contacts with long transfer lengths therefore require to be physically longer to remain efficient than those with short transfer lengths.

In the simplest case, the transfer length can be easily extracted from the basic TLM

described in Section 5.3 using the extrapolation of the linear fit for zero resistance as shown in Figure 5.3 [183]. In the case of alloyed ohmic contacts, the sheet resistance under the contact will be drastically different from that outwith the contact. In the case of non-annealed contacts, however, the sheet resistance might be expected to be similar in both regions. As a consequence, the transfer length is simply half the extrapolated length.

Transfer lengths were calculated using this method for the contacts fabricated for the third-generation 22 nm devices in Section 7.10. Average transfer lengths of 0.54 μm were extracted. As a result, it is to be expected that a 2 μm contact should behave similarly to a longer contact.

8.4.2 Thin metal recipes

Thinning the resist, as previously discussed, reduces the significance of electron scattering and hence increases resolution and contrast. As a consequence, the resist used for the definition of the short ohmic gaps should be as thin as possible. This has a clear consequence for the metal thicknesses that can be lifted off using the chosen resist. In addition, the use of thinned ohmic contacts, in conjunction with the etching of the silicon nitride pattern definition layer, should reduce the impact of the topography of the ohmic contacts on the subsequent definition of the delicate gate head.

Various gate metals were tested on c216 material for using TLM structures to characterise the resultant contact and sheet resistances.

Contact resistances for non-annealed contacts in principle should be dependent only on the composition of the bottom metal layer since the conduction band barriers are minimised by judicial doping control. As a result, two main variations on the recipe were attempted. Since the base metal layer used in previously-effective ohmic contacts was gold, 50 nm and 100 nm thicknesses of gold were deposited, in addition to a thinned version of the standard ohmic process.

	Thick	Thick(d)	Thin	Au-100	Au-100(h)	Au-50	Au-50(h)
$R_c/(\Omega.\text{mm})$	0.12	0.065	0.189	0.065	0.113	0.123	0.208
$R_{sh}/(\Omega/\text{sq})$	119.5	119.8	80.67	120.6	116.85	98.7	89.1

Table 8.1: Capped contact and sheet resistances measured using the Transmission Line Method on c216 for various metal recipes.

The standard "thick" recipe is shown alongside the results of Table 8.1 for reference. The 50 nm thinned Au/Ni/Ge recipe, which comprised 7 nm Au / 7 nm Ge / 7 nm Au / 8 nm Ni / 20nm Au, is labelled as "thin" whilst the two thicknesses of gold are "Au-100" and "Au-50" for the 100 nm and 50 nm contacts, respectively.

It was previously noted in Section 7.10 that the "thick" ohmic contacts measured at the end of a complete device process flow exhibited markedly improved contact resistances to the virgin contacts, annotated as "thick(d)". It seems likely that though the material is tailored for non-annealed contacts, there is some effect of the heat treatment during subsequent resist bake stages on the contact resistance, with potentially some additional germanium diffusion during processing. As a result, after an initial TLM measurement of the contacts, the samples were subjected to an overnight bake at 180°C in the resist oven, emulating a device process. The measurements were then repeated; these results are denoted by an (h) suffix.

It is noteable that 100 nm-thick gold indeed exhibits very low contact resistances on c216: lower than the standard recipe without heat treatment. The 50 nm contacts, however, displayed contact resistances around twice as large as the 100 nm contacts.

In addition, after heat treatment, both gold samples increased in contact resistance by 70-100 %. It would therefore appear that whatever the mechanism for current transport through the gold-only contacts, they are likely to be unsuitable for device applications. The thin contacts yielded relatively high contact resistances, around 57 % higher than the "thick" contacts, but are likely to improve with heat treatment.

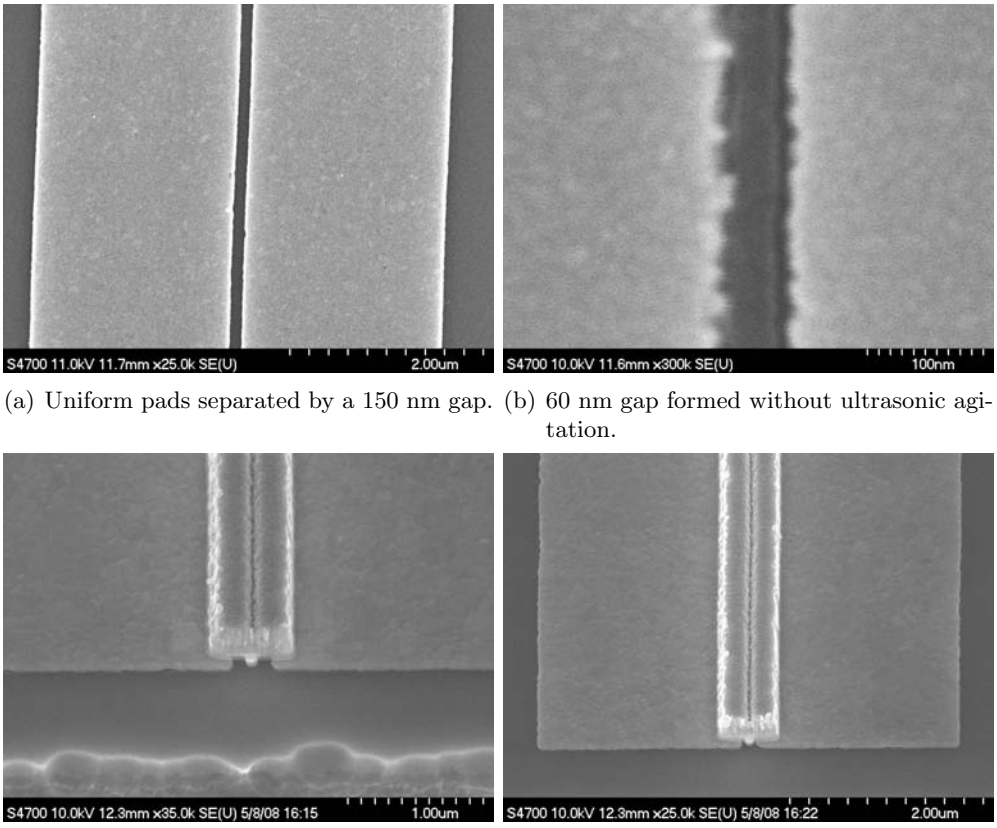
As a consequence, the Au/Ni/Ge recipes were considered more suitable for the fabrication of short ohmic gaps and the development focussed on the lift-off of 50 nm-thick pads.

8.4.3 Gap lithography

A pattern was designed with 2 μm pads separated by designed gaps from 1 μm to 25 nm. A large sample was coated with 50 nm silicon nitride, then a 4 % 2010 / 2.5% 2041 PMMA bilayer for lift-off, with a total thickness of approximately 140 nm, allowing metal layers with a maximum thickness of 100 nm to be patterned. The pattern was then dose tested using an 8 nA spot of 12 nm approximate diameter at 100 kV. The silicon nitride was then etched using the 4 m SF_6/N_2 RIE process and the "thin" 50 nm ohmic metal lifted off.

For an optimal dose of 220 μCcm^{-2}, gaps as small as 60 nm were realised without any additional processing for a designed size of 100 nm. By the additional undesirable use of ultrasonic agitation, gaps as small as 40 nm were realised. At this dose, the pads were fully cleared out, yielding well-lifted-off ohmic contacts with minimal edge flagging. The results are shown in Figures 8.9(a) and 8.9(b).

An upper gate level identical to that used for devices was then aligned into the gap between the pads to emulate the effect of further lithography.



(a) Uniform pads separated by a 150 nm gap.

(b) 60 nm gap formed without ultrasonic agitation.

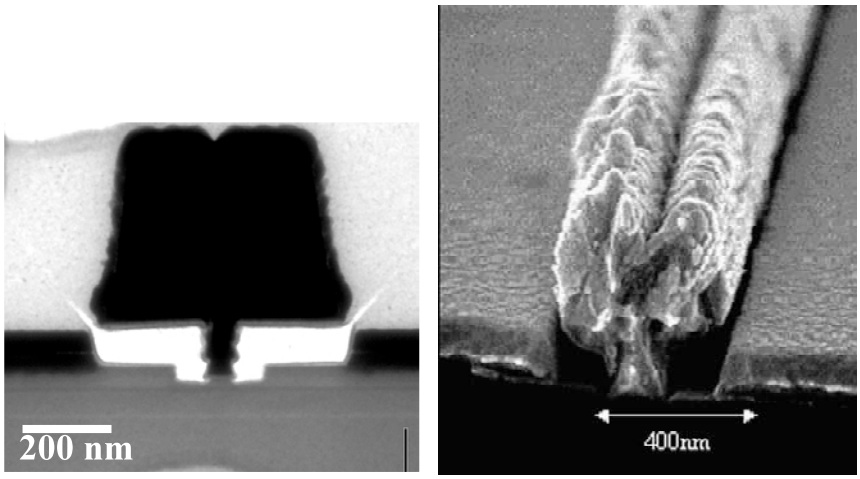
(c) 70 nm upper gate in 300 nm gap.

(d) 70 nm upper gate in 100 nm gap.

Figure 8.9: Short ohmic spacing resulting from new lithographic process.

As is clear from Figures 8.9(c) and 8.9(d), it is not only possible to fabricate the upper gate between the gaps, it is possible to form ohmic contacts well under the head of the T-gate, previously not achievable using self-aligned processes. Contrary to self-aligned methods, which produced 300 nm ohmic gaps by the use of the gate head as a shadow mask, there is a large separation between the ohmic pad and the gate structure at all points

of the device geometry, assuming thin contacts. A self-aligned structure necessitates the proximity of the ohmic pad to the gate head, resulting in very small, effectively random, nanometric separations caused by fluctuations in the evaporated films, shown in Figure 8.10.



(a) TEM section of a 50 nm self-aligned conventional HEMT.

(b) SEM section of a 120 nm self-aligned conventional HEMT.

Figure 8.10: Overview of conventional self-aligned HEMTs.

Using the two-step process, the ohmic contacts are separated from the gate by the silicon nitride spacer layer, which will have a clear dielectric effect on capacitance. The contacts, however, are fixed far from the gate head by the combination of the 100 nm-thick central section of the gate and the recession of the contacts into the silicon nitride film, whilst the gate remains above the film. As a consequence, there is increased vertical separation between the gate and the contacts. The combinatorial effect of the closer contact spacing and the presence of the silicon nitride dielectric, however, may have increased deleterious effects. As a consequence, the effect of such drastic decreases in contact spacing, particularly when placed under the gate head, is unknown and requires investigation.

The results achieved, together with the gate modules developed previously would allow the physical realisation of the idealised device structures of [107] and [22] used to extrapolate maximum device frequencies. There is therefore clear potential for the unlocking of additional device performance.

8.5 Material design

Chapter 7 clearly indicated the need for further work on the layer stack used for HEMT fabrication as a consequence of both insufficient scaling of the gate-channel separation and excessive parasitic resistances as a consequence of the cap doping.

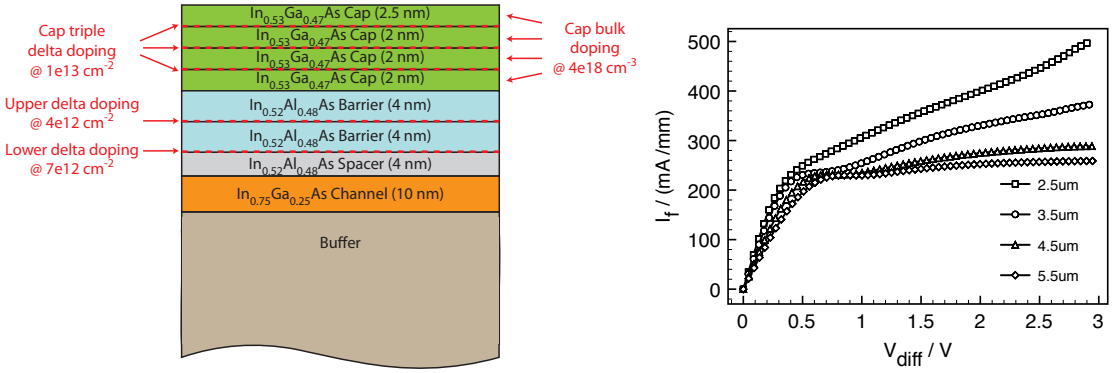
In addition, previous metamorphic material exhibited usual surface roughness as a consequence of the strain relief buffer. Although not intrinsically detrimental to performance, the surface roughness issues make reliable nanoscale lithography increasingly challenging as the feature sizes decrease. In particular, maintaining a consistent resist profile across a surface whose surface varies by up to 50 % of the critical device dimensions is undesirable.

As a result, several wafers were designed with decreasing gate-channel separation, all with 75 % InGaAs channels grown pseudomorphically on indium phosphide instead of relying on the metamorphic buffer on gallium arsenide. As a consequence, improved surface roughness uniformity would be expected.

Firstly, the cap layer structure was modified, with the aim of improving both contact resistances and sheet resistances to reduce source and drain resistances. Previous wafers, as outlined in Section 7.6, featured a triple-delta-doped 10 nm cap layer to provide the possibility of non-annealed contact formation by pinning the conduction band appropriately below the Fermi level. The cap thickness had proved suitable for short gate length fabrication, so the general approach was maintained.

In the new wafers, however, the cap was also background Si-doped at a doping concentration of $4 \times 10^{18}\text{cm}^3$. The reasoning behind this approach was to attempt to reduce lateral resistance in the cap, hence reducing the combinatorial parallel resistance of the cap and channel and lowering access resistances in the capped region. The most conservative wafer featuring the new cap structure was grown very similarly to c216, featuring a 13 nm gate-channel separation as opposed to the 15 nm separation used in c216. The layer structure is shown in Figure 8.11(a).

Problems were encountered in realising wafers with the new cap structure, however, as a consequence of the long times required to grow the complex structure of alternating highly background-doped InGaAs with layers of delta doping. Initial wafers featured recessed mobility of only around 7000 $\text{cm}^2/\text{Vs}^{-1}$ and sheet electron concentration of $6 \times 10^{11}\text{cm}^{-2}$, both well below the figures for c216. In addition, TLM I-V characteristics featured spurious non-linearity and suppression, similar to that seen during the surface



(a) c446 layer structure.

(b) Recessed TLM I-V characteristics for varying gaps.

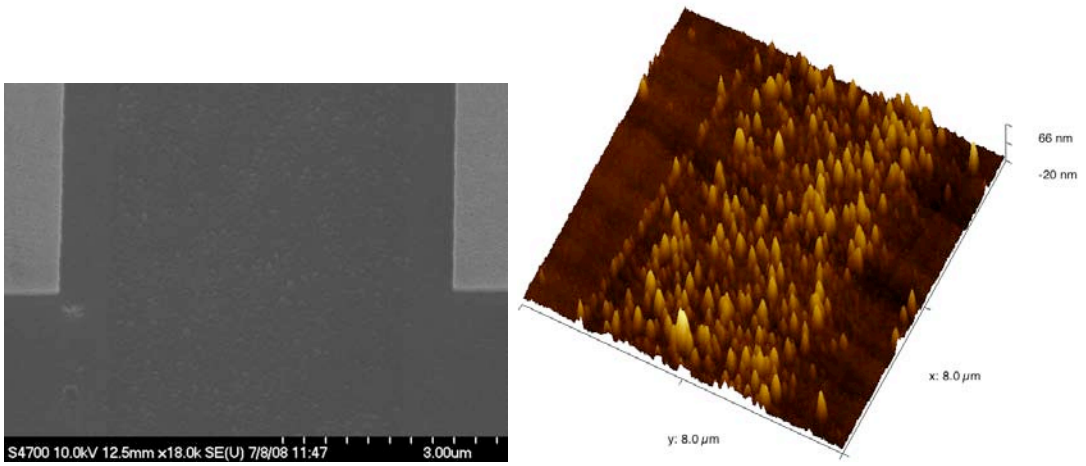
Figure 8.11: c446 layer structure and recessed TLM I-V characteristics.

treatment experiments of Section 7.9 and shown in Figure 8.11(b).

Additionally, in examining the surface of the completed c446 characterisation samples, extreme surface roughness was apparent in the etched area, as shown in the SEM and AFM images on Figure 8.12, with rms roughness on the order of 50 nm measured.

The effect of any damage resulting from the recess etching process was separated from the materials issues by extensive testing of the process parameter space on both the new wafer and c217, with particular effort expended on the effects of pH variation and etch time. In the case of c217, the effect of varying the pH balance of the etch chemistry was profound, with a fivefold increase in drive current sweeping the pH range from 5.3 to 5.9 as shown in Figure 8.13. Literature suggests the etch selectivity should vary with pH, with a decrease in etch rate with increasing pH [346]. All samples were additionally subjected to AFM analysis, and all were confirmed to have completely etched the InGaAs cap layer. As a consequence, the effects cannot be attributed to incomplete cap etching.

Prior devices were, by standard practice, etched using a pH of 5.5, which previously yielded acceptable etch selectivity. Though there was no change in the etch conditions, verified by the systematic replacement of each chemical component and measurement process, it is possible that an unidentified environmental shift caused an unexpected change in damage to the substrate at these conditions. In particular, redeposition or partial etching might be responsible at lower pH values. Under the new conditions, a pH of 5.9 was found to be preferable, with higher pH values etching the cap incompletely. The roughness effects additionally were not evident at a higher pH.



(a) SEM of recessed TLM gap. (b) AFM of recessed TLM gap.

Figure 8.12: Roughness apparent in all c446 recessed TLM structures. As previously, the recessed region is 1 μm smaller than the ohmic gap.

The prior range of post-recessing treatments was also repeated, and found to remain applicable.

In the case of c446, however, pH variation, though producing highly variable drive currents from the recessed TLM sites, was not recovered to acceptable values at any etch chemistry. In addition, the same problems were observed when recessing the new material using a selective citric acid process [237] and a "digital" approach [238, 239]. As a consequence, the succinic acid etch process itself was eliminated as the origin of the transport issues.

The issues were explained by consideration of active delta dopant in the new layers. Growing the new cap structure required the wafer to be subjected to elevated growth temperatures during a lengthy cap growth time, caused by the modulating cap doping. It was realised that this exposure could cause migration of the lower delta-doping planes used to dope the channel toward the surface, reducing effective dopant activation. This would explain not only the drop in electron density, but might also induce migration into the channel, reducing mobility.

A spread of Poisson-Schrödinger simulations was used to evaluate this suspicion using the Snider 1D solver previous employed.

The structure was simulated under four different surface potentials, designed to emulate

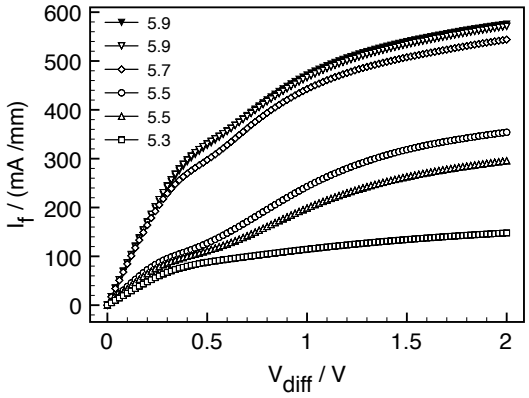


Figure 8.13: Effect of pH variation on 2.5 μm recessed TLM structures on c217.

any potential surface damage that might unexpectedly create additional surface states and affect Fermi level pinning. Within each of these surface potential scenarios, the doping efficiency was varied from the designed condition to a minimum of 50 % activation. In some cases, the upper delta doping plane was assumed to be unaffected, and the lower doping efficiency varied alone. Finally, the effect of a 1-2 nm shift in the delta dopant planes was simulated by shifting the position towards the surface by this distance.

Surface potential / eV	0.65	0.75	0.85	0.95
As designed	3.13×10^{12}	2.85×10^{12}	2.53×10^{12}	2.20×10^{12}
90% both delta	2.81×10^{12}	2.49×10^{12}	2.16×10^{12}	1.83×10^{12}
70% both delta	2.07×10^{12}	1.74×10^{12}	1.42×10^{12}	1.09×10^{12}
60% both delta	1.70×10^{12}	1.34×10^{12}	1.05×10^{12}	7.33×10^{12}
50% both delta	1.33×10^{12}	1.00×10^{12}	6.92×10^{11}	3.89×10^{11}
50% bottom delta only	1.73×10^{12}	1.41×10^{12}	1.08×10^{12}	7.62×10^{11}
50% bottom delta only + 1nm shift	1.37×10^{12}	1.05×10^{12}	7.33×10^{11}	4.28×10^{11}
50% both delta + 1nm shift	1.07×10^{12}	7.54×10^{11}	4.48×10^{11}	1.75×10^{11}
50% both delta + 2nm shift	8.17×10^{11}	5.07×10^{11}	2.24×10^{11}	3.52×10^{10}

Table 8.2: Simulated channel sheet electron density with various dopant activation efficiencies and position shifts towards the surface for a variety of surface potentials. All electron density figures have units of cm^{-2}.

As can be seen from the results of Table 8.2, several combinations of these effects are sufficient to cause the measured suppressed electron populations. In particular, a 50 % reduction in active dopant decreases the channel population by more than 50 %, whilst even a 1 nm shift in dopant position reduces the channel population by as much as 20-30 %. These effects are then exacerbated by any increase in surface potential incurred

during device processing.

As a consequence, the wafer growth conditions were altered to allow cap growth at a lower temperature, reducing dopant migration. The channel was then additionally grown at a hotter temperature to prevent its relaxation and increase its resistance to damage from the growth of the cap structures.

c577 wafer growth and further wafer design

The resultant wafer was then measured to have a mobility of 10500 cm^2/Vs^{-1} and a channel electron population of $1.72\times10^{12}cm^{-2}$, corresponding to 60 % activation of the total dopant from the simulations of Table 8.2. These are comparable figures to those of c216, which achieved figures of 11200 and $2.1\times10^{12}cm^{-2}$ respectively.

These figures are not unexpected for the reduced gate-channel separation and should indicate sufficient growth quality.

Ohmic contacts to this wafer were also characterised by TLM methods, yielding non-annealed contact resistances of 0.07 Ω.mm and sheet resistances of 128 Ω/sq using the standard thick ohmic recipe with no additional heat treatment. Contact resistances are therefore approximately half those of equivalent treatments on c216 (Table 7.3) with no major changes to channel transport whilst reducing the gate-channel separation to 13 nm from 15 nm.

Whilst c577 therefore represents an initial evolution of the wafer designs employed in early devices, the gate-channel separation remains large with respect to the gate length. As previously mentioned in Section 3.8.4, conventional wisdom suggests that a gate-channel separation of at least half that of the gate length is desirable [247, 273, 347–349], with reduced separations yielding enhanced transconductance and electron velocities.

A range of wafers with reducing gate-channel separations was designed, using Poisson-Schrödinger simulation with and without the cap structure in place to ensure conduction band compatibility with non-annealed processes and channel population with the cap removed.

Simulation results are shown in Figure 8.14, with the 13 nm separation representing the design for c577. The high cap doping and presence of a dopant plane close to the channel is sufficient, as previously discussed, to provide minimal conduction barriers to the channel,

Separation	13 nm (c577)	8 nm	7 nm	5 nm	4 nm
Cap	Identical for all, bulk doped 4×10^{18}, delta doped 1×10^{13} at 2 nm intervals				
Barrier	4 nm delta 4×10^{12} 4 nm delta 7×10^{12}	2 nm delta 1×10^{13} 2 nm delta 1×10^{13}	1 nm delta 1×10^{13} 2 ML delta 7×10^{12} 2 ML delta 1×10^{13}	2 ML delta 1×10^{13} 2 ML delta 1×10^{13} 2 ML delta 1×10^{13}	2 ML delta 1×10^{13} 2 ML delta 1×10^{13} 2 ML delta 1×10^{13}
Spacer	4 nm	3 nm	3 nm	2 nm	1 nm
Channel	Identical 10nm 75% for all				

Table 8.3: Overview of wafers of decreasing aspect ratio, where all doping has units of cm^{-2} and ML represents a single atomic monolayer, around 0.25 nm in thickness.

hence rendering material suitable for non-annealed contact formation. In the recessed case, however, Fermi level pinning of the barrier layer requires that the doping strategy be fundamentally altered with decreasing barrier thickness, as the surface pinning more directly impacts the conduction band geometry in thin barriers.

The 8 nm design comprises a 3 nm barrier and a double-doped 5 nm barrier at a maximum doping level of $1\times10^{13}cm^{-2}$ at each plane; doping sufficient to populate the channel with an electron density greater than that of the 13 nm case as a consequence of the reduced spacer thickness. Doping at higher densities, however, has proved previously to be futile, since the dopant activation does not continue to increase. The 8 nm design thus represents the most aggressively-scaled double-delta-doped case possible, and should also indicate the effect of a thinned spacer on channel transport, though many previous devices have used 3 nm spacers.

In order to further thin the barrier whilst maintaining channel population, triple delta doping of the barrier has been employed. The 7 nm design therefore represents the effect of this shift in doping strategy, with the spacer thickness unchanged from the 8 nm design. The 5 nm case represents the ultimate scaling of this approach, with minimally-thin layers between doping planes yielding a 2 nm barrier. Two monolayers has proved to be the minimum thickness reliably realisable by MBE, since a single monolayer may prove discontinuous. As a result of the thinned barrier, the 5 nm case also uses maximal doping in each plane. The 4 nm design therefore iterates off this approach, moving to a 2 nm spacer but maintaining a 2 nm barrier.

These wafer generations should therefore yield interesting results in short gate length

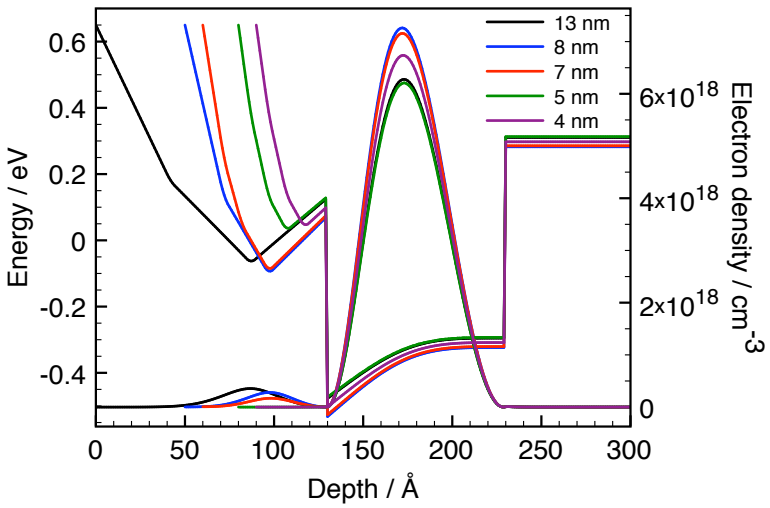


Figure 8.14: Comparison of conduction band profiles and electron densities for layer structures from 13 - 4 nm gate-channel separations. The surface has been offset such that the channels are concurrent.

devices. Topics expected to be of particular interest are firstly around the feasibility of realisation of these highly scaled structures, particularly regarding layer integrity and dopant activation and diffusion, since the designs rely on precise geometries and radical doping steps. Secondly, the performance of devices is of clear interest, firstly in comparison to simulated structures, but also in determining the methods of electron transport under various conditions. As outlined in Chapter 3, as the gate length is decreased and extrinsic effects minimised, one expects non-equilibrium effects to dominate HEMT transport. These structures should provide further opportunity for this, particularly in the determination of transport ballisticity, given the anticipated drop in mobility with reducing spacer thickness. Further scaling issues may additionally arise at separations approaching these scales, particularly gate-channel tunnelling.

Backdoping

The structures presented in Section 8.5 represent the minimum achievable gate-channel separations using top-doped channels whilst maintaining electron density. These reductions also occur at the expense of reduced spacer thickness, expected to lower mobility. As a consequence, performance may decline in such structures despite their interest as a metric of growth and transport.

A solution may be to abandon the usual top-only doping method used in Glasgow HEMTs and proved previously to optimise transport. By back-doping the channel, the spacer is moved to below the channel along with the dopant plane, leaving only the undoped barrier close to the surface. As a result, the barrier thickness can effectively be arbitrarily grown. As a side-effect, there is no reduction of the thickness of the energy barrier in the conduction band in the barrier as a consequence of doping, expected to reduce the probability of direct tunnelling from gate to channel.

The backdoping solution is therefore attractive, but requires abandoning the non-annealed strategy used in the gate-first methodology presented in this work, attractive for the simplification of high-yield precision lithography by smoothing of topography. A compromise might be found in the methods presented for defining ohmic contacts in the two-step process.

Since the ohmic contacts are evaporated into a trench etched in the silicon nitride, they are effectively recessed by the thickness of the silicon nitride, onto which resist is spun for further patterning. As a consequence, if the ohmic contacts are thinned to the same thickness as the silicon nitride, as for previous ohmic recipes used in the short ohmic spacing tests, the resultant surface is effectively still planar after ohmic patterning. An ohmic-first process can therefore be envisaged without compromising the gate lithography. The challenge in this case then becomes the realisation of suitable annealed ohmic contacts using very thin metallisations, given the thin layer of silicon nitride used in the 10 nm gate process.

Two initial backdoped wafers were designed by simulation, using identical 75 % channels as previously and a 3 nm spacer, giving 4 nm and 2 nm barrier thicknesses, shown in Table 8.4.

Separation	4 nm	2 nm
Cap	Bulk doped 4×10^{18}, delta doped 1×10^{13} at 2 nm intervals	
Barrier	4 nm	2 nm
Channel	10nm 75% as previously	
Spacer	3 nm	
Delta	8×10^{12}	

Table 8.4: Overview of 4 nm and 2 nm backdoped wafers. All doping has units of cm^{-2}.

The two wafer designs are identical apart from the barrier thickness, with identical doping giving the optimal results in each case. In the backdoping case, therefore, the role of

surface pinning appears to have reduced impact on the channel population. The Poisson-Schrödinger simulation results are shown in Figure 8.15.

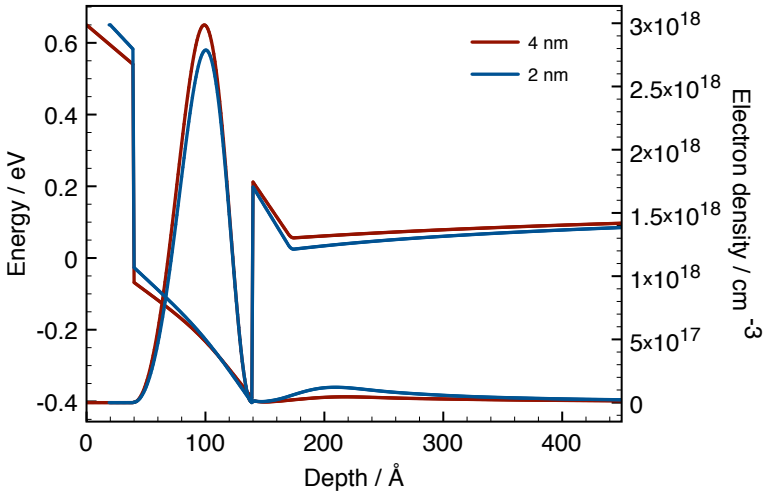


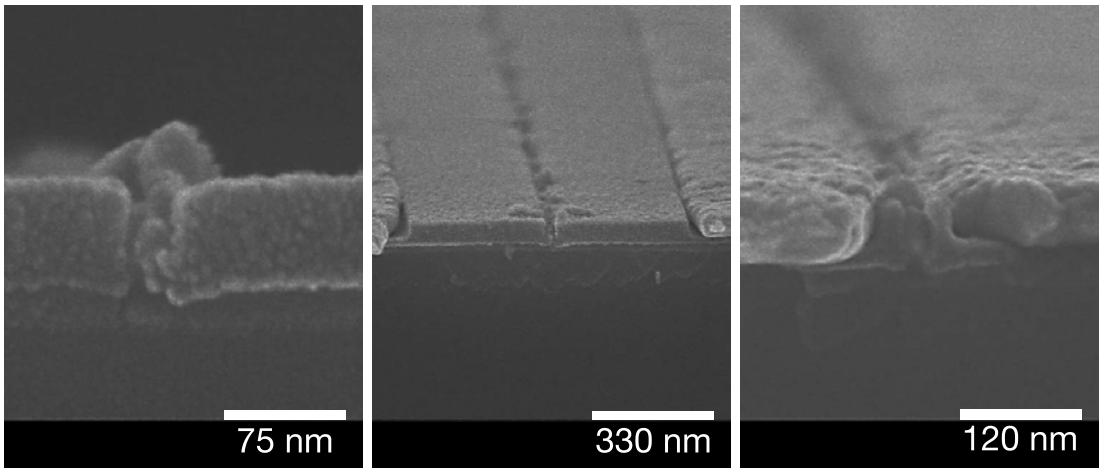
Figure 8.15: 1D simulation results comparing conduction band profiles and electron densities for 2 and 4 nm barrier thicknesses. The surface has been offset such that the channels are concurrent.

According to the simulations, the electron densities possible through backdoping techniques are considerably lower than the previously-designed structures, with the channel population approximately 50 % that of the multiply delta-doped structures of Figure 8.14. Backdoping, however, should prove to yield a mechanism for vertical transistor scaling for very short gate lengths whilst retaining high mobility, which may prove impossible with the more traditional structures.

In the case of highly scaled devices, there is likely to therefore exist an inherrent tradeoff in the InGaAs/InAlAs materials system. Epitaxial scaling of the existing structures is likely to preserve electron density, but at the expense of mobility, which in the equilibrium case, will reduce electron velocity. Migration to a backdoped strategy would likely yield high mobility, but reduced electron density. The overall high frequency performance of highly-scaled devices built on the two epitaxial layouts would likely be determined by the role of non-equilibrium effects as the gate length is scaled.

8.6 Device fabrication

The new alignment and ohmic processes were used to fabricate 22 nm devices on c216 in the same manner as previously. The gate foot was defined, then ohmic contacts of source-drain separations from 60 nm - 1 μm aligned using the Penrose alignment techniques. On the first set of devices, an unexpected variation in topography was noticed, and the I-V characteristics of a TLM structure measured to establish the effectiveness of the ohmic contacts. The measurements revealed a problem with the contacts, most probably incomplete etching of the silicon nitride. Test samples were therefore cleaved for cross-sectional SEM analysis. The results extracted are shown in Figure 8.16.



(a) 22 nm gate in 500 nm ohmic gap, showing nitride filling of recess trench.

(b) Overview of 22 nm device with 500 nm source-drain spacing.

(c) 22 nm device with 60 nm source-drain spacing.

Figure 8.16: Under-etched short ohmic gap 22 nm devices.

As is clear from the images of Figures 8.16(b) and 8.16(c), the SF_6/N_2 RIE etch has indeed terminated before etch completion, with a 5-10 nm thickness of silicon nitride remaining under the ohmic contacts. Clearly, this would prevent device operation. The images extracted, however, are useful in verifying several pertinent details.

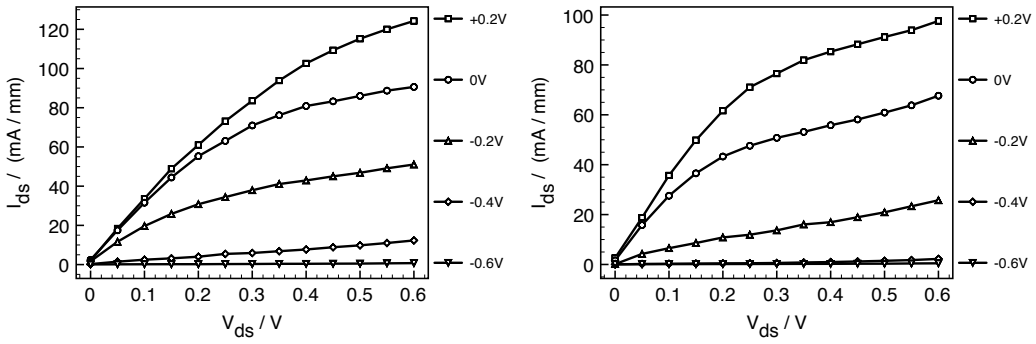
Firstly, the images show clearly the recess formation under the silicon nitride, with conformal deposition of the film over the sloped edges of the etch trench. This is particularly clear from Figure 8.16(a). The contact edges additionally meet intimately with the silicon nitride through the complete film thickness. The silicon nitride indeed completely encapsulates the active region, even at very small source-drain separations.

Additionally, Figure 8.16(c) clearly shows both the radical geometries that can be achieved using this ohmic process and the excellent alignment accuracy possible from correlation-based alignment, with the gate perfectly centred in both the recess trench, evident from the slope of the silicon nitride, and the ohmic pads themselves.

Interestingly, the 60 nm case, whilst perfectly realised, may prove to be beyond the current limits of practical HEMT fabrication. Firstly, the recess trench, realised by wet etch, may prove to be the limiting step, since the recess length limits the potential proximity of the ohmic contacts and recesses shorter than 60 nm have proved difficult to reliably fabricate, as described in Section 7.10. Additionally, the current upper gate level features a lower geometry of 60-70 nm. As a result, shorter ohmic separations will result in source-gate-drain shorting. Potential misalignments increase the probability of the pads contacting the recess or gate.

As a consequence, though the process enables the fabrication of ohmic gaps as small as 60 nm, future devices were designed to feature a minimum gap dimension of 100 nm.

22 nm devices were again fabricated on c216 using the process, using the intended 4 minute ohmic etch time, yielding functional devices with ohmic spacing ranging from 1 μm to 100 nm. Measured device characteristics, however, showed seriously suppressed drain currents as is clear from Figure 8.17. It should be noted that 22 nm devices with ohmic contacts fabricated using both the 50 nm-thick and conventional ohmic recipes exhibited the same issues.



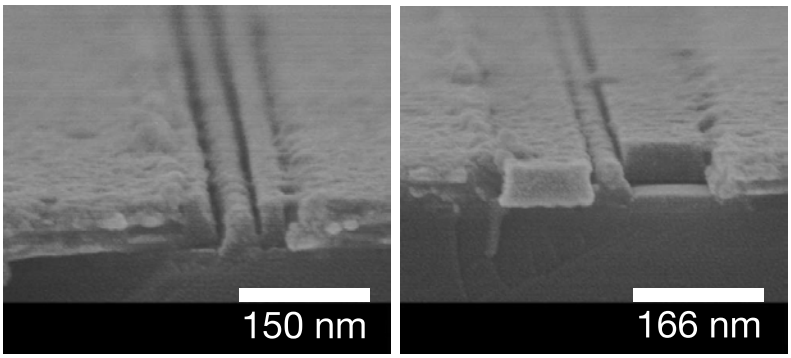
(a) 800 nm ohmic separation.

(b) 100 nm ohmic separation.

Figure 8.17: V_{ds}/I_{ds} characteristics of devices with 800 nm and 100 nm ohmic separations.

Capped TLM measurements likewise displayed suppressed currents, with contact resistances of 1.1 Ω/mm and sheet resistances of 180 Ω/sq resulting. It therefore appears

that there is a mechanism at work causing both resistances to increase. This could in part be due to under-etching as previously, though this is not evident from Figure 8.18. Additionally, the dry etch completion was verified by AFM prior to ohmic metal evaporation.



(a) 100 nm separation. (b) 200 nm separation.

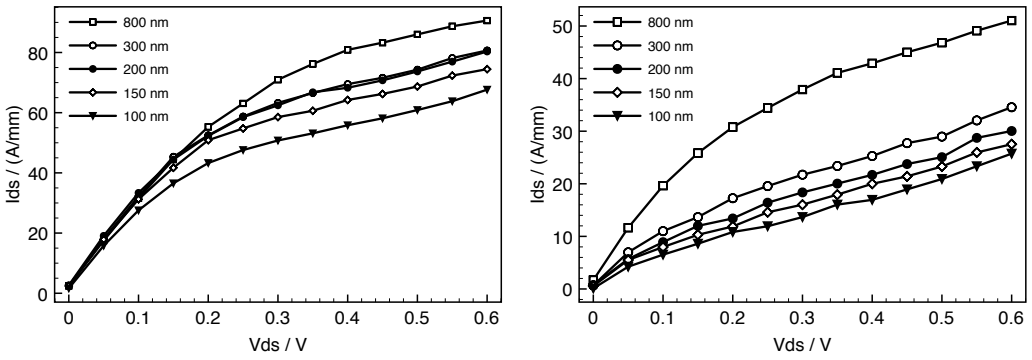
Figure 8.18: SEM cross-section of measured short ohmic separation devices.

All devices over the 1 μm - 100 nm range of ohmic separations, however, showed effective transistor action over the measurement range, with full pinch-off and no major kinks in the I-V characteristics. It should be noted that applied drain voltages were limited to 0.6 V to avoid breakdown which would preclude further measurements. The characteristics beyond this voltage are therefore as yet unknown.

Interestingly, and contrary to expectations, however, devices with a decreasing separation exhibited decreasing output currents, as is clear from the comparisons of Figure 8.19.

This is unexpected, since the device access resistances should decrease as the area of the ohmic gap decreases. As these access resistances decrease, one intuitively expects the output currents to increase, with the intrinsic region under the gate contributing in greater measure. Additionally, however, devices with shorter ohmic separations pinch off more rapidly, with devices with longer separations maintaining increasingly larger currents as gate bias is increased. This verifies the reduction of the parasitic resistances, since extrinsic transconductance would be expected to increase given reduced source resistance (Equation 3.40).

To understand this effect, the s-parameters of the structures were measured, allowing the parasitic components to be extracted comparatively. Devices were therefore measured from 0-67 GHz as previously. To extract the parasitics without affecting the intrinsic



(a) $V_g = 0V$. (b) $V_g = -0.2V$.

Figure 8.19: Comparison of output characteristics of short-ohmic devices for various applied drain bias.

components, which could conceivably also vary between devices of varying ohmic separations, cold measurement techniques were employed.

Cold FET measurements were first proposed by Dambrine, et al. [188] as a technique for extracting FET parasitics by measuring devices under zero applied drain bias. By transforming the s-parameters to z- and y-parameters, the various extrinsic circuit elements can be manipulated at each stage of transformation.

Measurements at zero applied drain bias therefore allow extrinsic resistances and inductances to be extracted. The capacitances are more complicated, since the channel conductivity interferes with the overall measured capacitance. If the device is biased such that the channel is pinched-off, the effect of the channel should be negligible in comparison to the parasitic capacitances.

The devices were measured at both conditions, and the resistances extracted. The results were plotted versus contact separation in Figure 8.20 and a linear fit made to the data.

As expected, the extracted resistances decrease rapidly with source-drain separation. The results, however, are not as anticipated, with results for a 1 μm separation on the order of 120 Ω, rather than the 6.5 Ω expected for a 50 μm-wide device. As expected, therefore, source-drain resistances are over an order of magnitude greater than expected for the gap dimension, explaining the order of magnitude reduction in drive currents.

As is clear from the plot, whilst the evolution of extracted resistance with separation is

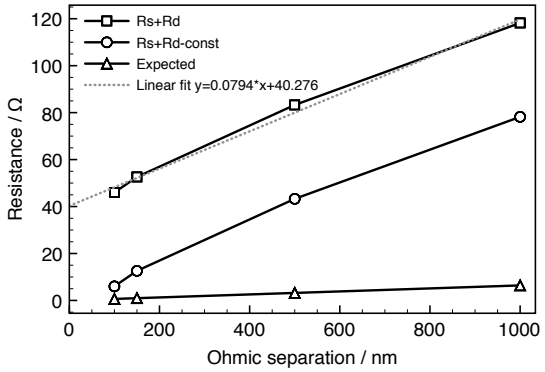


Figure 8.20: Extracted resistances for various source-drain separations.

linear, the extrapolation does not intercept zero resistance at zero separation, as would be expected given idealised contacts on material of uniform sheet resistance. Instead, there is an offset of around 40 Ω, which is not unexpected given the anticipated high contact resistance. By removing this offset from the data points, a more physically expected value is extracted, where the extrapolation to zero separation does result in zero axial resistance.

A further effect is, however, noticeable. Figure 8.20 plots the expected total resistance for various separations, neglecting contact resistance and basing the calculation on recessed sheet resistance figures extracted previously for the wafer. The theoretical plot features resistances greater than an order of magnitude less than the extracted figures for the material. The current reductions therefore cannot be attributed only to etch residue, since the effective sheet resistances are grossly increased for all separations, though the effect does appear linear. There consequently appears to be a process at work which greatly increases the sheet resistance of the device material between the ohmic contacts.

Given that the resistances are linear with separation, whilst capacitances increase slightly with decreasing separation, the extracted parasitic values do not intrinsically explain the unexpectedly proportional relationship between measured drive currents and contact separation of Figure 8.19. It would therefore appear that the effect at work additionally affects channel transport without further increases to the parasitic resistances.

Further 22 nm devices were also fabricated on c577, the latest material of 13 nm gate-channel separation outlined in Section 8.5. These devices exhibited identical problems.

These devices therefore have three main issues:

- Extremely high contact resistance.
- Extremely high sheet resistance.
- Additional mechanism decreases drive currents for decreasing separations.

One possibility is that there might be some edge effects at work, perhaps due to carrier redistribution corresponding to variations in Fermi level pinning around the ohmic contacts or gate. Effects of this sort could potentially deplete the channel, whilst the axial parasitic resistances would remain dominated by the high cap sheet resistance. An explanation of this sort may also be married to specific processing issues, with, for instance, increased etch residue for shorter gaps.

A specific issue is that the RIE tools underwent major maintenance before the fabrication of these devices, with a resultant need for recalibration of the etch rate for all structures. In particular, a repeat of the experiments of Figure 7.12 was required to determine the new etch time appropriate for 22 nm gates, producing a revised gate etch time of 5m 30s to reproduce the pattern transfer results previously achieved. Since the surface remains exposed to the etch chemistry for minimal times, it was hoped that the effects on the exposed surfaces would be minimal, as previously.

It is, however, possible that these changes have additionally caused the etch process to become more damaging than previously. Damage of this sort, affecting all dry-etched regions, could explain all the effects observed in these devices, since both the gate and ohmic regions would be affected. It would be expected, therefore, that devices of decreasing source-drain separation would exhibit increased damage to the channel, since an decreasing region would remain covered by the silicon nitride. If channel resistivity is increased along a larger percentage of the channel, electron velocity will decrease given the lower injection velocities and reduced acceleration.

10 nm devices were also fabricated on c577 in parallel with the 22 nm devices, using the high-resolution method of 10 nm gate fabrication of Section 8.3.1. These devices, in addition to the problems outlined, exhibited minimal gate control as shown in Figure 8.21.

It is at this stage unclear as to whether the poor control over the channel is due to underetching of the gate brought about by the revised etch parameters, or a loss of control

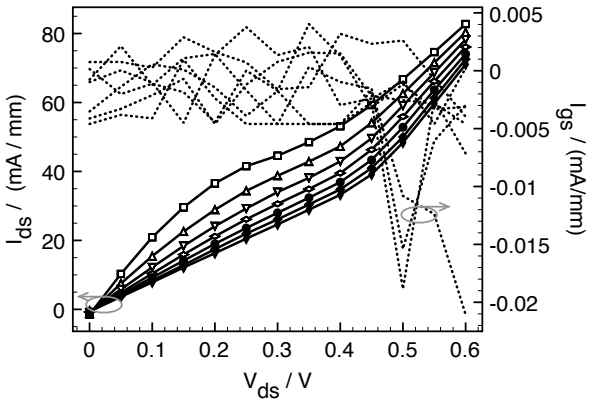


Figure 8.21: $(I_{ds}, I_{gs})/(V_{ds}, V_{gs})$ characteristics of a 10 nm, 25 μm-wide device with 200 nm ohmic separation fabricated on c577.

as a result of the improper lateral scaling shown in Figure 3.27. It is probable that this very sensitive process now requires proportionally longer over-etch under the revised silicon nitride etch process. This is further reinforced by the low gate leakage observed in Figure 8.21, though the gate leakage trends over the limited drain bias range indicate there is some path for leakage through the silicon nitride. The etch may therefore only be partially complete, giving minimal control over the channel. It is, however, noteworthy that the detrimental effects of poor scaling would be expected to be relevant in these devices, since the gate-channel separation is much larger than the gate length.

Further SEM and TEM are required to confirm these observations, in addition to further work to re-evaluate the effects of relevant plasma processing following the changes in the etch conditions.

8.7 Summary

This chapter has detailed several developments to the processes proposed in Chapter 7. These processes include two methods for scaling the gate pattern transfer processes to 10 nm, using either high-resolution lithographic and pattern transfer processes, or by length reduction processes employing the conformal deposition and anisotropic etching of silicon nitride.

Methods for realising ohmic separations as small as 60 nm have been verified and integrated with all the processes proposed in this thesis for the fabrication of highly scaled

devices.

Several designs for the realisation of highly scaled epitaxial structures have been simulated, showing the potential for excellent performance with very small gate-channel separations and addressing the issues observed with the wafers grown for the needs of Chapter 7. The initial challenges in the growth of the most relaxed of these new structures have additionally been investigated and circumvented in preparation for the realisation of more aggressively-scaled designs. In particular, the need for the investigation of the limits of double-delta-doped structures and the potential move to backdoping has been highlighted.

Functional devices with source-drain separations as small as 100 nm have been fabricated, integrating all the processes developed in this thesis. Further problems were, however, encountered with apparent damage in these devices, potentially as a result of changes in the RIE conditions. 10 nm devices were fabricated, but also suffered problems, though the exact cause may be a combination of processing issues and intrinsic scaling limitations. There is therefore a requirement for both continued investigation into plasma damage under the revised etch conditions and further microscopy into the 10 nm device structures.

The resolution of these issues and continued work on epitaxial development is expected to yield high performance, fully-scaled devices at gate lengths as small as 10 nm as a result of this work.

9. Implant Isolation

9.1 The motivation for implant isolation

As the gate length of transistors is reduced, a further effect which begins to play an important role for narrow devices in particular is the capacitance arising from the feeds connecting the gate fingers from the active device region to the gate bondpads or waveguides. As the gate capacitances scale with the gate length, the additional parasitic component from the feed contribution becomes increasingly significant. Conventionally, this feed must connect the short and delicate gate across the isolation mesa sidewall. In order to traverse the vertical distance of 50-100 nm, large metallised gate feeds are required to maintain good contact and must overlap onto the active region. This overlap on active material gives rise to substantial capacitive and sidewall leakage effects which can contribute significantly to the degradation of device performance [103].

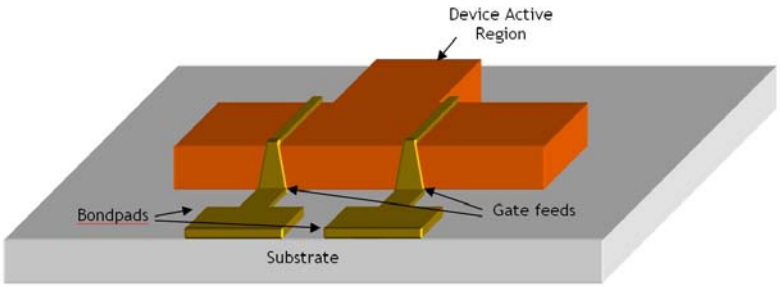


Figure 9.1: Conventional FET mesa isolation, showing the gate feed overlap.

One solution is to eliminate the problem by eliminating the mesa itself, planarising the device and removing the need for a large overlapping metallisation. Such an approach can also greatly simplify lithography, which is strongly affected by topographic non-uniformity of the substrate.

Implant isolation provides a method of reducing the conductivity of a semiconductor by damaging the lattice, such that the effect is similar to removing the material altogether, as would be the case for conventional mesa etching techniques. By bombarding the material to be used for the device with energetic ions, the crystal lattice can be altered, displacing atoms and creating trapping which reduces conductivity [350]. If this effect can be well controlled, the areas exposed would be expected to be effectively insulating, isolating the active region from the surrounding material.

This implant isolation mechanism has been successfully used with III-V semiconductors for many years, but has been predominantly limited to gallium arsenide [351, 352], due to the ease of isolation by use of proton or other light ion bombardment. In the case of indium phosphide and related materials, however, isolation is far more difficult, since the smaller bandgap and higher intrinsic carrier concentration of the material reflect a lower intrinsic resistivity [353], reasoning which indeed underlies its frequent use in high-speed electronic and optoelectronic devices.

9.2 Theory

Implantation-induced isolation can result from two possible physical processes. Firstly, isolation may occur as a result of damage produced in the lattice by the physical interaction of the incident ions on impact with the semiconductor. This physical interaction may take one of two forms: nuclear collisions where the ion strikes the nucleus of a lattice atom, or electronic, whereby the incident ion collides with bound electrons [350].

As an ion enters the semiconductor, electronic stopping effects cause the loss of energy and slow the ion. Once substantially slowed, the ion has a high probability of colliding with lattice nuclei. On collision with a lattice atom, energy is transferred as momentum is conserved, which may cause a recoil deflection of the struck atom. As a result of a combination of these physical collisions and Coulombic effects between the ion and atomic nuclei and electrons, atoms can be displaced entirely from the lattice, creating interstitial atoms unbound to the surrounding lattice and corresponding lattice vacancies at the sites of the dislocations [350]. Both ions and atoms are likely to collide further with the semiconductor lattice, and can cause further atomic dislocations: a recoil cascade. This tends to occur towards the latter part of the progress of an ion into the semiconductor, leaving many defect vacancy sites towards the terminal range of the ion.

These lattice defects are then present throughout the implanted region of the semicon-

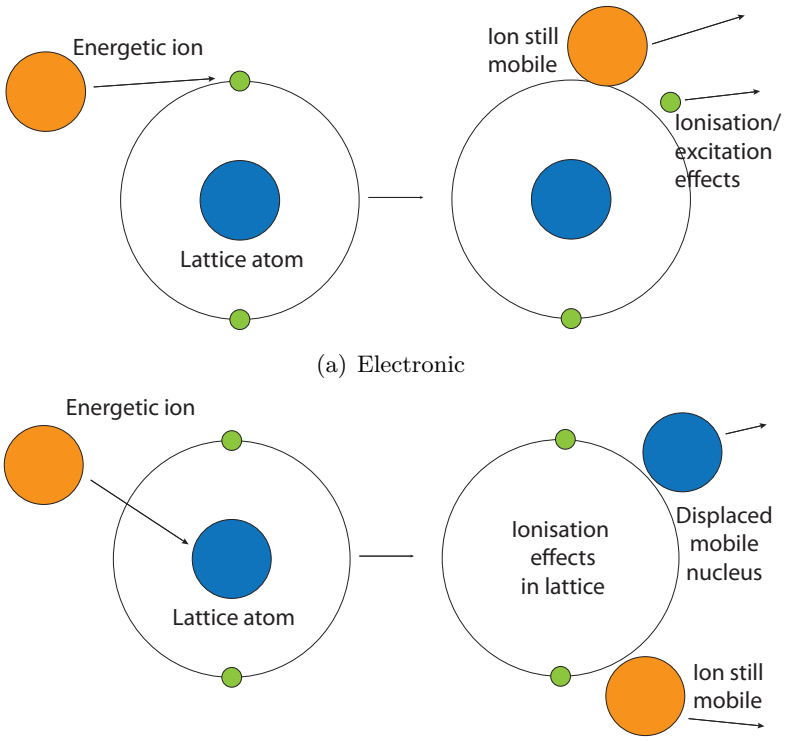
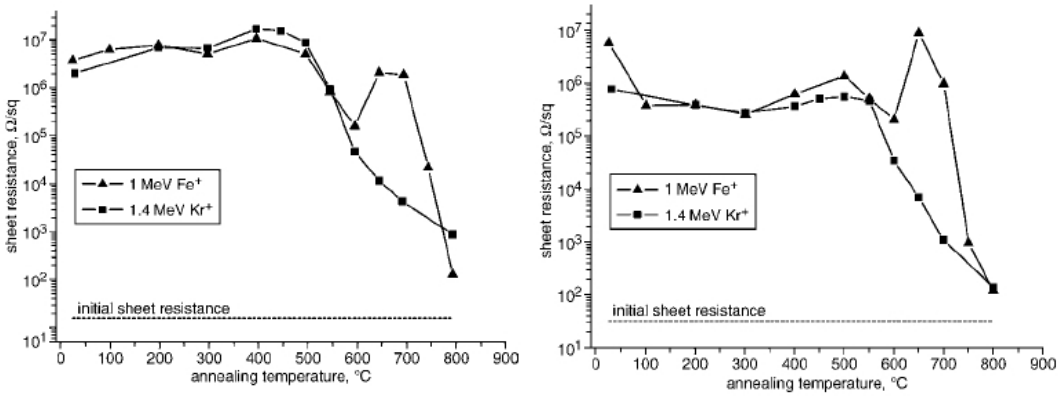


Figure 9.2: Collision processes comprising the physical mechanisms of implant isolation.

ductor, causing deep-level trapping which results in the degradation of the conductivity of the material. This is known as damage-induced isolation.

The second method by which isolation may be achieved is by the creation of chemically-active deep levels as a result of nuclear interactions at the lattice level. Chemical isolation results from the implantation of an ion species which will interact with the lattice in such a way that bonding may occur between the ion and the material, its dopants or impurities. A requirement for chemical isolation is that the ion implanted must be able to substitute into a lattice vacancy created during a nuclear collision. Once substituted, the new atom forms a part of the lattice, and changes its chemical properties. The key to chemically-induced isolation in n-type semiconductors is to implant a species which will result in the creation of deep acceptor levels or complex trapping areas which serve to trap mobile electrons [354]. As a result of the requirement for a substitution reaction, an annealing step may be essential to provide the necessary energy for bonding and promote the formation of new trapping complexes [350].

For indium phosphide and its compounds, unimpressive results have been shown for isolation for most ions. $In_xGa_{1-x}As$ in particular seems to be a challenge to isolate, given its extremely narrow bandgap and high intrinsic electron concentration, with resistivities of around $10^5\Omega$/sq reported [353]. From a device perspective, poor isolation will result in large leakage currents, detrimental to performance, with conductivity through the InGaAs channel of greatest concern. Unsatisfactory isolation will also lead to greatly increased waveguide loss, hampering circuit performance.



(a) InP (b) $In_{0.53}Ga_{0.47}As$

Figure 9.3: Substrate resistivity following Fe and Kr implants as a function of annealing temperature. After Too, 2004 [355].

Results from the University of Surrey [355], however, indicate that it appears to be possible to achieve good isolation over a wide temperature range by the use of either damage or chemically-induced isolation, with the achievement of resistivities of around $10^7\Omega$/sq, both in InP and $In_{0.53}Ga_{0.47}As$, shown in graphs reproduced in Figure 9.3. Damage-induced isolation has been investigated by the implantation of krypton ions, whilst chemically-induced isolation has been achieved by the substitutional implantation of iron.

9.3 Experimental Design and Sample Preparation

As a result of these promising results, experiments were undertaken to investigate the possibility of using krypton and iron isolation implants as a realistic and viable isolation method for planar III-V devices. These experiments were designed to attempt such implants on a real HEMT device layer stack.

Since the layer stack consists of considerable thicknesses of $In_{0.48}Al_{0.52}As$, the effect of damage-induced isolation for the Kr^+ implant is difficult to predict; however, the cap and channel layers in which we are most interested are both comprised of InGaAs, although for the high indium content $In_{0.75}Ga_{0.25}As$ channel desired here, the layer may behave more like InAs.

Given the large proportions of InGaAs in the device stack, the iron implant process with the high-temperature annealing step was also attempted, both to monitor the effect on the sheet resistance as a result of the implantation and the effects on the stability of the epilayers with these high temperatures. Since MBE generally occurs at around $450 - 500°C$, temperatures higher than this may cause significant elemental migration as well as undue thermally-induced stress on the epitaxial layers. The annealing routines described, however, last for only two minutes, so would possibly be sufficiently undamaging to device epitaxy. Iron implants with an annealing phase have therefore also been investigated for the devices in question.

The implants were to be carried out on MBE-grown metamorphic InGaAs/InAlAs HEMT layers on a GaAs substrate, wafer C080, similar in design to the layers designed for the short gate length processes developed during this project. The most relevant active device layers for isolation are the device channel and the highly-doped cap, as can be seen in Figure 9.4. It was therefore necessary to implant only as deep as the epitaxially-grown device layers, around 50nm, much less than the range of hundreds of nanometres required by Too et al [355] in their initial work. Simulation was therefore undertaken to analyse the required optimal energies for effective isolation of the device stack.

Implants have been tailored using James Ziegler's SRIM (Stopping and Range of Ions in Matter) Monte Carlo simulator [356] to produce implantation solutions for both ions which will result in as uniform damage as possible to the active device depth, with a particular focus on the channel and heavily-doped cap.

For both ions, it was necessary to use multiple implant techniques, since for the HEMT layers in question, there is a need for relatively constant damage over a fairly shallow range: virtually impossible to achieve with a single implant. Multiple implant techniques allow the combination of several damage profiles to give a uniform spread over the complete depth by combining implants in order of decreasing energy. In SRIM, it is possible to analyse the vacancies created as a result of simulated ion collisions and their recoil cascades, which, for damage-induced isolation, is a good method of modelling lattice resistivity.

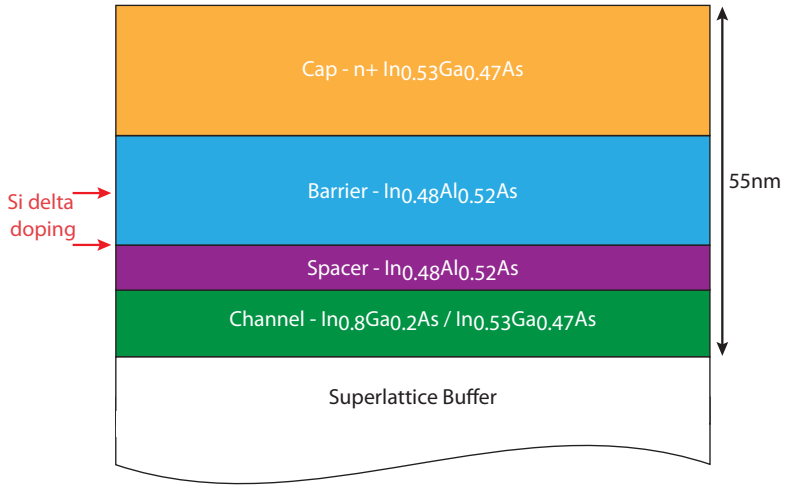


Figure 9.4: Double delta-doped metamorphic device layer stack based on a 7.5nm $In_{0.8}Ga_{0.2}As$ / 7.5nm $In_{0.53}Ga_{0.47}As$ composite channel.

For the krypton isolation, a 150keV / 10keV double implant technique was chosen, using a 1:1 fluence ratio between the two energies, which appears to give good uniformity across the key layers, as can be seen in Figure 9.5.

The ultimate goal of chemically-induced isolation, as in the case of the Fe^+ implantation, is to look not just at the locations of the vacancies created, since these will provide only damage-related trapping data, but also at the final depths of the implanted ions, since these will interact with the lattice to provide what will become deep trapping centres on annealing. Consequently, it is important to look at both plots, shown in Figure 9.6.

A 100 keV / 9.2 keV double implant was selected for the iron implant at a fluence ratio of 20:3, since this combination gives both a reasonable damage profile and a good ion distribution through the active layers. An even damage profile is desirable for the iron implant as the chemical isolation effects cannot become a factor until annealed at high temperature. Since the epitaxy of the devices may not survive this annealing, it is also desirable to investigate the damage-induced isolation as the samples are annealed and chemical isolation becomes a factor.

On further consultation with the University of Surrey, these conditions were slightly modified to 150 / 25 keV for krypton at doses of 1×10^{13} / $2.1 \times 10^{12}cm^{-2}$, and 100 / 15 keV at doses of 1.6×10^{13} / $2.7 \times 10^{12}cm^{-2}$ for iron, in order to balance damage between samples and allow accurate comparison. Split samples were also used, such that the effect

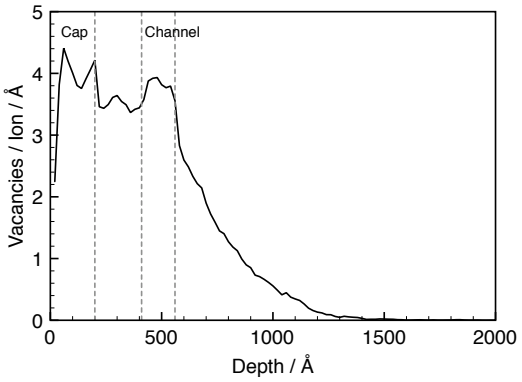
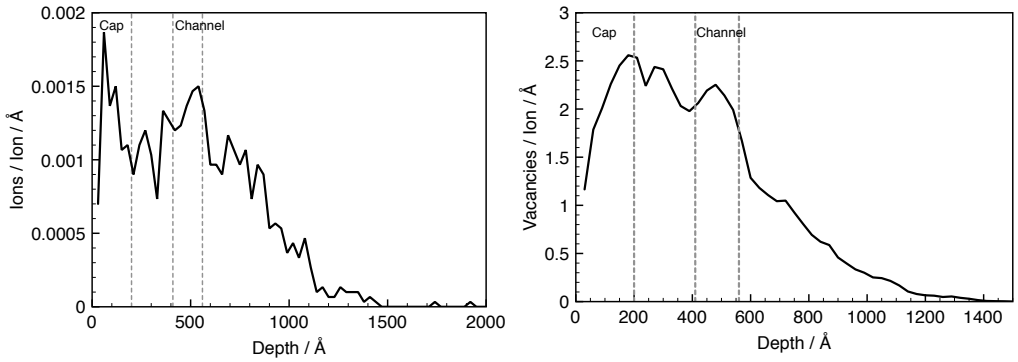


Figure 9.5: Vacancy creation resulting from Kr^+ double implant at 150/15keV.



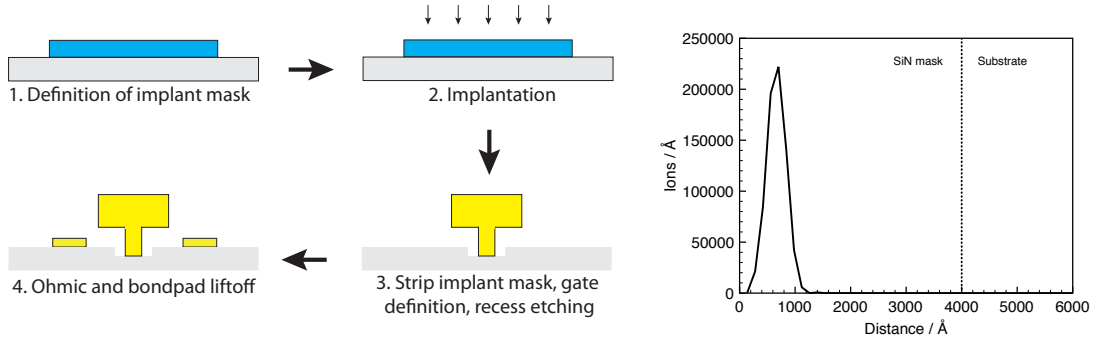
(a) Ion Distribution (b) Damage Distribution

Figure 9.6: Ion and damage distributions following a 100keV/9.2keV double implant with Fe^+.

of the second implant could be determined by direct comparison to a singly-implanted sample.

The implant process is integrated into the device process flow by first patterning markers only onto the substrates to be used for the devices, allowing subsequent levels to be aligned. Silicon nitride hard masking was used to specify the areas to be exposed to the implant, and was uniformly deposited across the substrates, before alignment of a masking level to the markers and dry etching the excess nitride. Further SRIM simulation (Figure 9.7(b))showed that 400 nm silicon nitride was adequate to stop any penetration of implanted ions in masked regions, whilst electron beam exposure of thick Sumitomo NEB-31 negative tone resist was used to form the hard mask. The SF_6/N_2 process used

for short gate length definition was additionally used here to etch the silicon nitride, yielding a vertical profile. Both deposition and etch processes were previously shown to be low-damage in Section 7.4.2.



(a) Implantation process flow (b) Ion stopping in hard mask

Figure 9.7: Hard masking processes for implant isolation. The mask is defined in 400 nm silicon nitride.

The samples were then implanted as described and returned for further processing. An ohmic contact layer was deposited, allowing the electrical properties of the material to be characterised by van der Pauw and transmission line methods. Gates and bondpad levels were then subsequently deposited. The process flow is outlined in Figure 9.7(a).

The non-annealed ohmic contact process described in Section 4.6 was used. This allowed the complete processing to be undertaken at low temperatures and the temperature dependence of the iron implants to be analysed without complication.

9.4 Measurements and Results

The electrical properties of the material were analysed initially by capped and recessed van der Pauw methods, using standard four-point test structures. Measurements were taken for all four samples for both implanted and unimplanted material, both to measure the effectiveness of the resultant isolation and any damage to the surrounding active material.

Since the resistivity of the implanted material was unusually high for this measurement system, the bias current used for the measurements had to be greatly reduced, and some difficulty was experienced in the measurement process. As a result, sheet resistance mea-

surements were also taken by means of the TLM (Transmission Line Method). Both capped and recessed vdP and TLM structures were measured. With regard to isolation implants in particular, the use of both capped and recessed structures allows the effectiveness of the implant at channel depth to be measured separately from that at the cap level, closer to the surface, allowing comparison between single and double implant strategies. These measurements also allow detection of any damage or reduction in mobility in the unimplanted surrounding material.

9.4.1 Non-annealed TLM Measurements

Measurements were performed on all four samples prior to any annealing studies, and show a direct comparison of results between capped and uncapped structures for single and double implant strategies for damage-induced isolation. The results are outlined in Table 9.1

Sample	Capped / (Ω/sq)	Recessed (Ω/sq)
Kr1 - 150 keV	8.19×10^5	1.46×10^6
Kr2 - 150/25 keV	9.42×10^5	1.52×10^6
Fe1 - 100/15 keV	7.36×10^5	1.33×10^6
Fe2 - 100 keV	6.85×10^5	1.18×10^6

Table 9.1: TLM Transport data for implanted material

Firstly, it would appear from the results in Table 9.1 that the double implants produce slightly higher resistivity than their single-implant counterparts, especially for capped krypton-implanted material. This would suggest that the higher energy implants, at least for krypton, do not cause damage all the way to the surface.

Secondly, it would appear by comparison of all the structures in Table 9.1 that the recessed structures have an intrinsically higher resistivity than the capped, implying that all the implants have been much less effective at shallow depths. This is somewhat surprising for the double implant, which was designed to be damaging to the surface layers. This may be due to the use of an overly high energy for the second implant, resulting in fewer ions stopping and causing vacancies in the shallow surface regions.

The overall sheet resistance for the channel alone, extracted from the recessed measurements, however, is dramatically higher than the pre-implanted value, and comparable or superior to those found by Too et al [355] for both krypton and iron implants. The values

shown for iron are based purely on damage-related isolation, not from any deep level-related chemical effects. The iron implant without anneal appears to produce similar sheet resistance values to those expected, since the initial high-resistance peak associated with an unannealed iron implant is annealed out by resist baking steps incurred in further patterning steps post implant. It is worth noting that these values are associated with the 80% indium InGaAs channel, and consequently this result could be seen as slightly surprising given the small bandgap of InAs.

9.4.2 Non-annealed van der Pauw Measurements

Van der Pauw measurements were also carried out in order to measure the material transport properties of unimplanted surrounding material, in case of any detrimental effects as a result of straggle or insufficient masking. For completeness, measurements of implanted material were also carried out, though sheet resistance values found by this method did not correlate well with TLM-extracted measurements, the vdP measurements being on average some 60% lower than those from the TLM. This has been attributed to a loss of measurement accuracy associated with the use of the smallest possible bias current and extremely high resistances.

	Implanted			Masked		
Sample	R_{sh}	uH	n_{sh}	R_{sh})	uH/	n_{sh}
Kr1 - 150 keV	345489	0	3.66×10^{14}	110.234	6636	8.54×10^{12}
Kr2 - 150/25 keV	294954	0	4.10×10^{13}	98.661	5686	1.11×10^{13}
Fe1 - 100/15 keV	242376	0	2.35×10^{14}	95.244	6019	1.09×10^{13}
Fe2 - 100 keV	257114	0	2.84×10^{14}	81.694	5142	1.49×10^{13}

Table 9.2: Van der Pauw transport data for implanted and unimplanted material. R_{sh} is sheet resistance with units of Ω/sq, uH is mobility with units cm^2/Vs, n_{sh} is sheet electron density, units of cm^{-2}

.

As can be seen from Table 9.2, the masked regions retain a very high sheet carrier concentration of around $1 \times 10^{13} cm^{-2}$ and mobilities in the range of 5000-6000 $cm^2/(Vs)$, with sheet resistances around 100Ω/sq. These values are as expected for capped van der Pauw measurements for this material system, and so it would appear that the implant process has not resulted in any undesirable damage to the surrounding material.

9.4.3 Iron Implant Annealing Studies

Since the highest resistivities should occur in InGaAs for iron implants at highly elevated rapid anneals of an optimal 650-700 °C [355], it was decided to attempt to anneal an iron sample at both a lower temperature and at the 650 °C target temperature. Since the MBE growth occurs at around 500 °C, it was of interest to discover the temperatures at which the epitaxial layers begin to break down, even with a rapid (60 second) anneal and temperature ramp. The measurements previously discussed used non-annealed ohmic contacts and have therefore undergone no heat treatment other than standard 180 °C resist bake steps.

Since the singly-implanted sample displayed the worse resistivity, it was decided to subject it to the two annealing steps whilst retaining the double implant sample for further work.

(a) Implanted Material

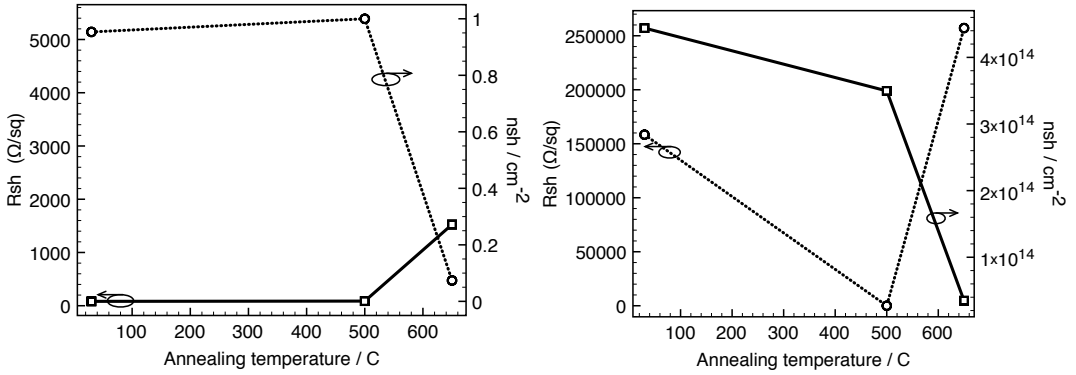
	Capped			Recessed		
Sample	R_{sh}	uH	n_{sh}	R_{sh})	uH/	n_{sh}
Non-annealed	257114	0	2.84×10^{14}			
500°C	198937	1	2.72×10^{13}	201820	4	7.31×10^{12}
650°C	4683	30	4.44×10^{14}	5809	232	4.63×10^{12}

(b) Masked Material

	Capped			Recessed		
Sample	R_{sh}	uH	n_{sh}	R_{sh})	uH/	n_{sh}
Non-annealed	81.694	5142	1.49×10^{13}			
500°C	86.717	5389	1.34×10^{13}			
650°C	1525	474	8.65×10^{12}	3933	314	5.06×10^{12}

Table 9.3: Van der Pauw transport data for annealing studies on singly-implanted capped and recessed structures. Symbols and units are as previously.

From the measurements, it can be seen that as the sample is annealed at 500 °C, the sheet carrier concentration in the implanted areas drops markedly, accompanied by a slight increase in mobility. This is contrary to the original results after Too et al [355], which reported a moderate increase in resistivity at a 500 °C annealing temperature. In unimplanted regions, there is negligible change in transport properties, which would imply there has not yet been significant degradation in the quality of the epitaxy. For comparison purposes, the original van der Pauw data has been used at the same bias current as previously.



(a) Masked (b) Implanted

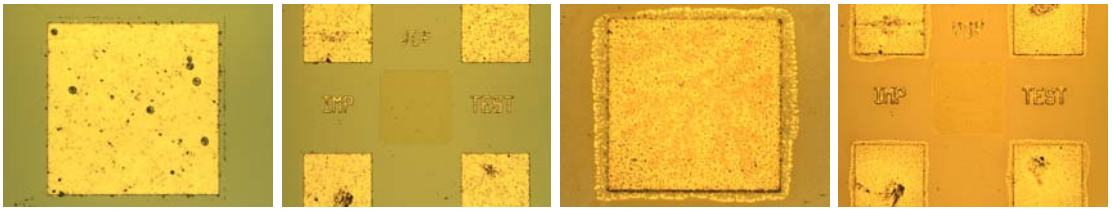
Figure 9.8: Mobility and sheet resistance of iron-implanted samples with annealing temperature.

In the implanted case, resistivity in fact decreases at 500 °C although sheet electron concentration drops. It seems logical that, although the activation anneal may activate some trapping centres, as predicted by Too, et al., there is also a thermally-related healing process which acts to anneal out the implant damage, though electron density may decrease. It therefore appears, since the net resistivity drops at elevated annealing temperatures, that the dominant damage mechanism in the iron-implanted epitaxial structures, as for krypton implants, is damage. There appear, therefore, to be few advantages to pursuing an iron-based chemical isolation process over a damage-based heavy ion implant such as krypton.

On raising the annealing temperature to 650 °C, which should provide optimal isolation according to reference [355], it was clear that the epitaxial layers had been badly damaged. Sheet resistivity increased almost twenty-fold in the unimplanted material, accompanied by a tenfold drop in mobility. For comparison, recessed measurements were also made for samples annealed at this high temperature, showing channel mobility both to have dropped significantly in unimplanted regions, but also to have effectively equalised between implanted and unimplanted areas.

In effect, the epitaxial layer structure has been so damaged that the intermixing of lattice components has completely destroyed the isolation characteristics of the unmasked regions and the heterojunction interface which provides the excellent transport qualities of a HEMT.

In addition, the ohmic contacts were also not optimised for annealing at such elevated temperatures, and were badly degraded, as can be seen from Figure 9.9, with sputtered metal shorting any narrow gaps. Van der Pauw measurements, however, use widely spaced contacts, and should not have been significantly affected. This degradation, however, would be too severe to produce functional devices, so were the epitaxial problems solved, a new ohmic recipe would be required.



(a) 500 °C (b) 500 °C vdP structure (c) 650 °C (d) 650 °C vdP structure

Figure 9.9: Ohmic contact degradation at high annealing temperatures.

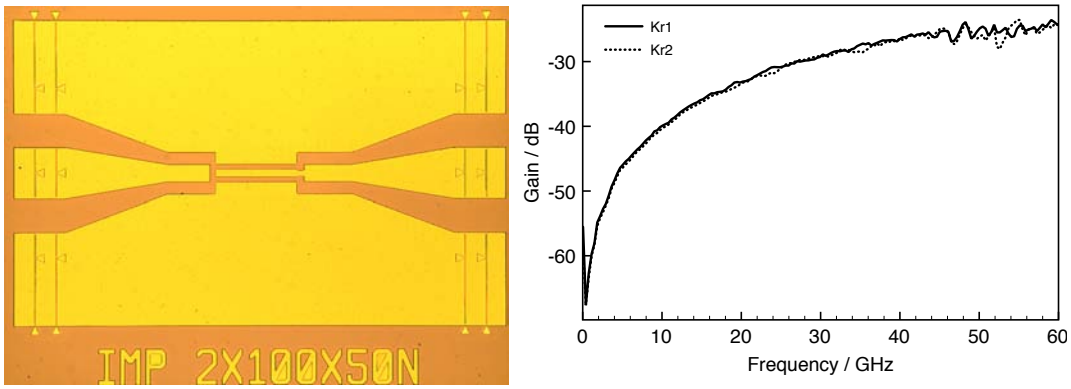
9.4.4 RF Loss Measurements

By forming a "dummy" structure by depositing coplanar waveguide bondpads in the conventional device layout described in Section 4.6 and shown in Figure 9.10(a), but excluding any of the active device structures and forming the device on implanted and therefore insulating material, it is possible to perform RF analysis on the implanted material to measure its transmission characteristics for use in active devices.

S-parameter measurements were taken using a Wiltron Vector Network Analyser from 0-60 GHz. By looking at the magnitude of S21, the forward transmission gain, it is possible to determine the leakage occurring across the frequency band through the implanted material. Well-isolated material should produce minimal attenuation of the input signal across the whole band, which should be clear from the magnitude of S21.

Conventional devices are formed on an etched mesa, so the passive areas are etched down to the semi-insulating GaAs, yielding high waveguide transmission at all measurement frequencies. The implanted material, however, produces resistivity an order of magnitude lower than mesa isolated structures as is clear from Table 9.1, and larger leakage currents than conventionally-isolated structures. As a result, one would expect reduced attenuation on implanted material in comparison to the conventional case.

As can be seen from the S21 magnitudes shown in Figure 9.10, there is more than 20



(a) Dummy FET structure

(b) Kr Implant S21 Loss

Figure 9.10: "Dummy" FET structures and transmission loss on double krypton-implanted samples.

dB attenuation across the band for both singly and doubly-implanted samples, despite slightly lower resistivities than had been hoped for. At low frequencies below 10 GHz, attenuation is greater than 40 dB, whilst attenuation drops as frequency increases, as for conventionally-isolated devices. The drop in attenuation appears to level off at around 40 GHz, remaining at around 25 dB up to the measurement system limit of 60 GHz. This loss profile mirrors almost exactly that of a conventionally-isolated device, albeit with a 5 dB drop in attenuation across the complete frequency band.

Interestingly, there is virtually no distinction between the two implant samples, whilst resistivity measurements differ. Whilst the double implant sample does damage the cap layer to some extent, the implant has not been as successful as desired in the cap, which may explain the correlation with these measurements.

9.4.5 Transmission Line Measurements

RF measurements were also taken for coplanar waveguide transmission lines fabricated on isolated material, in order to compare transmission characteristics and relative loss for the implanted material with the semi-insulating GaAs which forms the substrate for transmission lines on conventionally-isolated MMICs. Implant-isolated devices imply complete circuits fabricated on implanted material, so it is important to ensure losses are not too great to make practical MMICs.

As a result, CPW transmission lines were fabricated at lengths of 100, 200, 500 and

1000 μm and characterised across the 0-60 GHz frequency sweep. This allows a quantitative extraction of loss through the semi-insulating implanted substrate in terms of signal attenuation per unit length, which may be compared to data from existing mesa-isolated material.

Looking to the Kr-implanted samples, a change of less than 0.25 dB was observed across the complete 60 GHz spectrum for the 100 and 500 μm waveguides. Looking to all structures, we can see that the loss appears, as would be expected, to be linearly increasing with line length and increasing with frequency. For all measurements, the calibration was poor at frequencies greater than 40 GHz, but it is clear that the general trend continues to higher frequencies.

By taking account of all measurements for each sample and normalising, it is possible to extract a loss measure per unit length. Since this is frequency-dependent, it can be seen graphed in Figure 9.11. From the graph, it is obvious that the single krypton implant appears to show slightly reduced loss as compared to the doubly-implanted sample, and both some 2 dB/cm greater than a mesa-isolated structure at low frequencies, rising to around 4 dB/cm around 40 GHz.

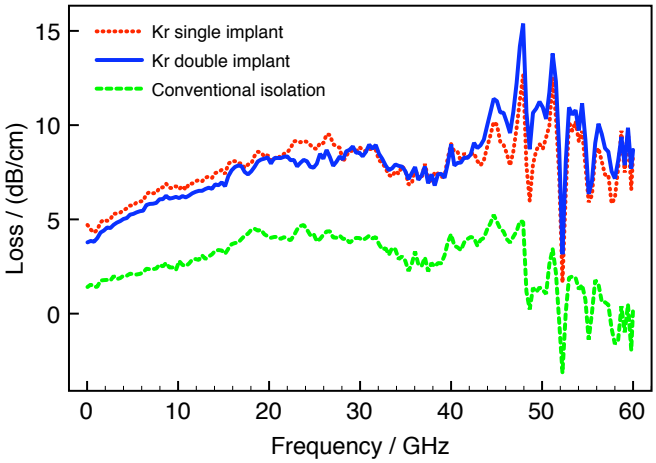


Figure 9.11: Normalised loss comparison for implanted and conventionally-isolated CPW structures.

In effect, this result shows that for relatively short lines, there will be a very slightly greater attenuation for implanted samples, whilst longer lines of over 1 cm could show an increased attenuation of around 35-40% at high frequencies. This may be acceptable for some applications, or could be combined with some secondary mesa isolation to de-

crease loss whilst retaining planar device characteristics. This may improve the parasitic contributions of these planar devices, whilst improving long-run isolation using a mesa approach.

Unfortunately, there were recessing problems during gate definition such that the first batch of fabricated devices would not pinch off satisfactorily. This was found to be due to incomplete or "patchy" etching of the silicon nitride masking used during implantation, confirmed using a Nomarski optical interference contrast microscopy technique.

These results were, however, promising, and would imply that the krypton isolation system may well be suitable for the fabrication of implant-isolated planar InGaAs/InAlAs devices and circuits.

9.4.6 Implanted device fabrication

In order to improve the isolation achieved using the Kr process, further samples were implanted, varying several different implant parameters. The shallow implant energy, firstly, was reduced to the originally designed 15 kV energy in order to deduce the impact of the changes made during the first implant. Secondly, a spread of implant fluences was used in order to determine the effect of potential hopping conduction as a result of excess implant dose.

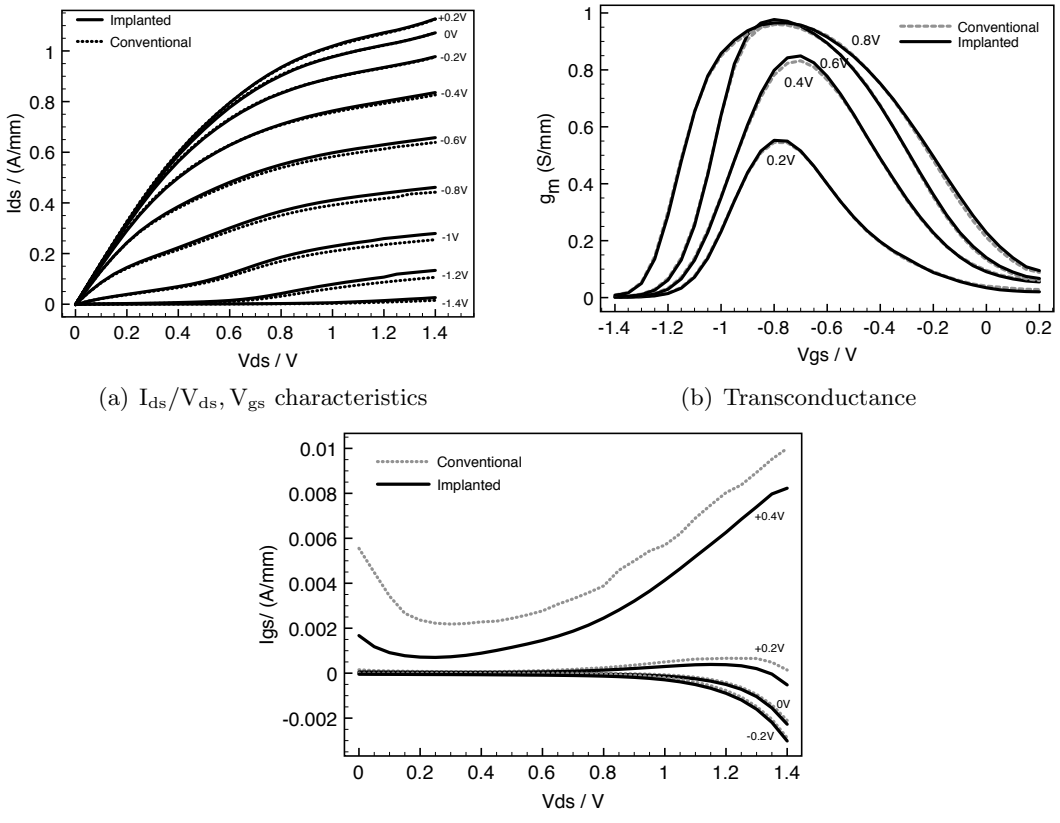
The samples were processed as previously, using identical hard masking and device fabrication processes to the first samples.

TLM measurements were again used to measure the resistivity of the implanted regions. Averaging across several sites, as previously, the measured transport data were markedly improved over the previous sample. In implanted regions, the resistivity was 1.15 MΩ/sq according to capped measurements, and 2.19 MΩ/sq in recessed regions. This is an increase of 17% in capped regions and 44% in recessed samples.

Conventionally-isolated devices were also fabricated on the same sample to allow accurate comparison. These devices were fabricated with the gate feeds entirely off-mesa. Although this dramatically decreases gate yield in conventionally-isolated devices, it also allows a direct comparison between the conventionally-isolated and implanted devices to analyse any damage arising from the use of an implant process. The merit of the implant step in this scenario would effectively be to increase the gate yield with the feed remaining off-mesa.

Markers were fabricated prior to implantation, then the hard mask defined, the sample implanted, then the mask removed using a blanket SF_6 etch. Conventionally-isolated structures were then formed on the sample. 60 nm T-gates were then defined on both implanted and conventionally-isolated sites using the single step PMMA/LOR/UVIII process described in Section 4.6 and the gates recessed using a single 45 s succinic acid recess. Conventional ohmic contacts and bondpad structures completed the fabrication process.

The devices were measured both at d.c. and r.f. from 1-110 GHz.



(a) $I_{ds}/V_{ds}, V_{gs}$ characteristics

(b) Transconductance

(c) Gate leakage

Figure 9.12: Comparison of conventionally-isolated and double krypton-implanted 60 nm 100 μm-wide device performance.

As is clear from Figure 9.12, the output characteristics of the implanted and conventional devices are virtually identical, both exhibiting peak drain currents in excess of 1.1 A/mm. In addition, both sets of devices yielded peak transconductances of close to 1 S/mm.

These output characteristics are virtually identical between the two sets of devices.

A small difference in performance was noted in the measurements of gate leakage, where at positive bias, the implanted devices yielded lower leakage currents, as is clear from Figure 9.12(c). Small improvements were also noticeable at negative gate bias. The improvements are not unexpected: conventional devices have a large metal feed contacting the channel directly at the mesa edge, contributing significantly to leakage. Implanted devices do not feature gate metal across the mesa, hence should yield improved performance.

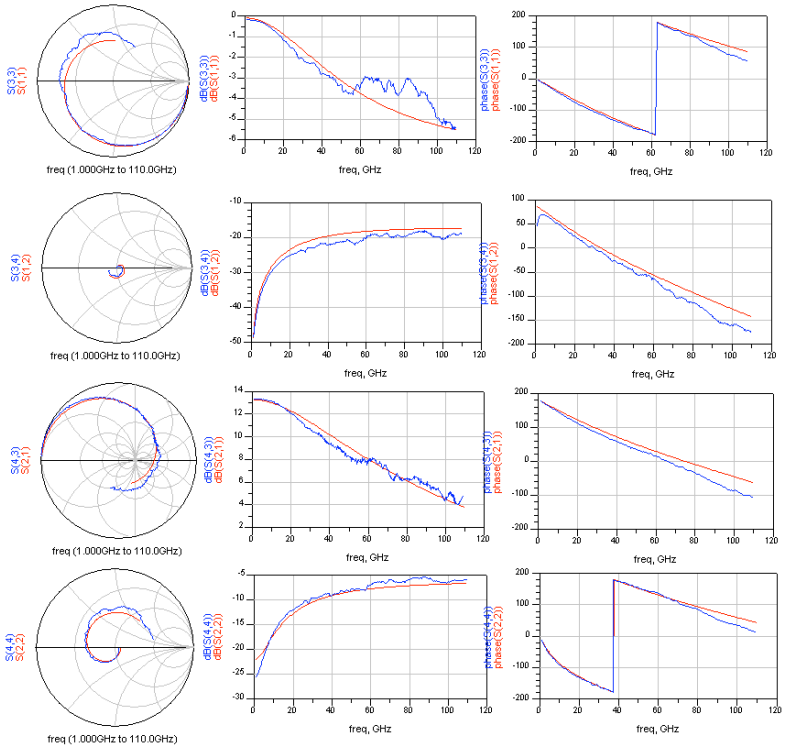


Figure 9.13: S-parameter fit of equivalent circuit model for 100 μm -wide 60 nm implanted device.

The s-parameter measurements were extracted as previously (Figure 9.13, and the equivalent model built using physical parameters as before. The circuit was then de-embedded from the coplanar waveguides and h21 and MSG/MAG were graphed, allowing the cutoff and maximum frequency of oscillation to be extracted.

The equivalent implanted and conventional devices exhibited identical s-parameters and both had identical figures of merit: an extracted cutoff frequency of 420 GHz and a

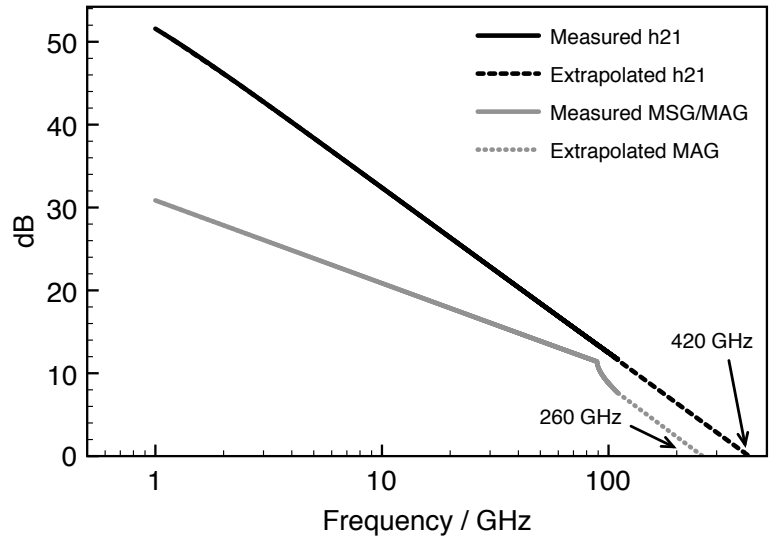


Figure 9.14: Measured and extrapolated h21 and MSG/MAG, showing extracted f_t and f_{max}.

maximum frequency of oscillation of 260 GHz.

These are respectable figures of merit given the implanted nature of the devices and the 60 nm gate length: the extracted f_t is only slightly lower than the best expected from conventional 50 nm metamorphic devices [357]. It would therefore appear that the implantation process has not induced any significant deterioration in device characteristics. In combination with the acceptable loss measurements extracted from the CPW lines, it would appear that the double 150 kV / 15 kV Kr^+ implant may provide a route to realise acceptable planar implant-isolated devices and MMICs.

9.5 Implantation intermixing effects

One of the side-effects of implant isolation is caused by the straggling effect of ions as they are implanted, which causes ions to travel laterally within the semiconductor as opposed to the ideal vertical profile. Implanted ions tend to follow a 3-dimensional Gaussian profile as they are implanted, with large lateral travel towards the end of the ionic range [172, 353].

The effect of this is that any masking strategy will result in some ion displacement under the edges of the mask, unexpectedly implanting some areas. This will result in the

damage of a small amount of material intended to remain active. For heavy ions, this effect is compounded by the large recoil cascades discussed previously, as the deflection of recoiling atoms will not be as uniform as the initial implant, and so increased travel under the mask results.

Due to this effect, the epitaxial structures will become intermixed around the edges of the theoretically unimplanted active region, resulting, for example, in the formation of some InGaAlAs quaternary in the boundary region between InGaAs and InAlAs layers. Although this effect is unlikely to cause problems for typical wide devices, as devices are narrowed, particularly for low-power or digital applications, these "smeared" regions may begin to comprise an appreciable proportion of the active region.

An interesting method of analysing this phenomenon is to simulate the effect of implantation around a point and log the co-ordinates of each collision for both ions and recoiling atoms, as is possible in SRIM. By tracking the final collision in a cascade, an estimate of the final resting place of atoms can be determined to a short distance. An excellent example is to track the displacement of silicon atoms from the virtually two-dimensional delta doping planes during the implant process. These co-ordinates can then be plotted in 3D to give a visual representation of the possible displacement effects which may occur.

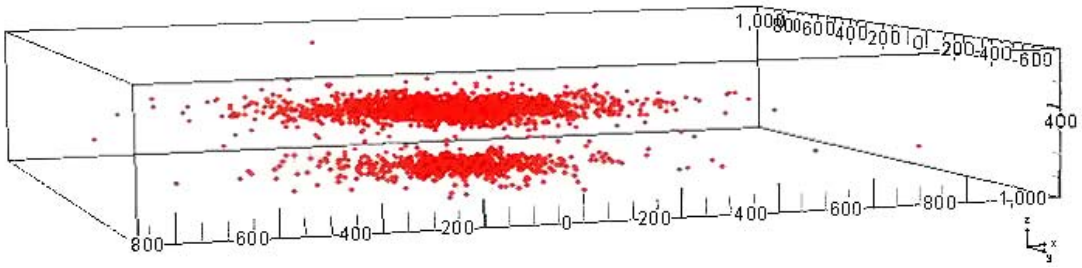


Figure 9.15: Displacement of silicon atoms in a double delta-doped device, modelled in SRIM by 150 keV Kr^{+}implantation through point (0,0,0). Displacements of up to 100 nm result. Axis units in angströms.

As can be seen from Figure 9.15, considerable displacements can result in all three axes as a result of implantation through a single point. Of particular interest are those atoms which travel outwith their original layer depths and also travel laterally, since it is these atoms which will reach an eventual resting place in a layer in which they were not intended. This simulation would appear to confirm that this effect may prove to play a role in very narrow devices. The formation of quaternary compounds around the edges will produce slightly different bandgaps, and these regions may even be lightly doped by

the displacement of silicon from the delta doping layers. This will degrade the conduction band profiles and hence electron concentrations around the edges of the active region, which may prove to have an effect on confinement effects and device performance.

In order to determine the reality of these effects, it is desirable to measure the resistivity characteristics of implanted material to determine if the simulated dopant straggle effects cause an appreciable deterioration in real implanted material. To this end, several masking structures have been designed. Each of these structures consists of two pads at a fixed separation, connected by a narrow region which should remain unimplanted. By placing ohmic contacts on the pads, resistance between the contacts can be measured. Several of these structures have been designed with varying widths of connection region. Neglecting these smearing effects, resistance should be linear with width. If, on measurement, non-linear effects are observed, then dopant smearing will have played a role.

It is noteworthy that a similar effect will also occur in conventionally-isolated material, since surface depletion will occur in any unpassivated semiconductor due to the presence of surface states, discussed in Section 3.5. For increasingly narrow areas, this depletion will comprise a higher proportion of the active region.

Lines of various lengths and fixed 300 μm width were realised in the silicon nitride hard mask, then implanted using the double krypton implant process to investigate the effects of intermixing on the measured resistance. Ohmic pads and all other processing were as for previously-described samples. The lines were then measured using four-probe I-V techniques, and the linear resistance extracted from each. The results are shown in Table 9.4.

The large-area lines were used to calculate the sheet resistance of the material, which, at 118 Ω/sq, correlate fairly well with the extracted van der Pauw measurements of Table 9.2. It is noteworthy that these figures are extracted using different techniques, and the low-current van der Pauw method previously used yielded inaccurate results as noted in prior results. Sheet resistances were also calculated for each fabricated geometry and corresponding resistance. As is clear from Table 9.4, the calculated sheet resistance remains fairly constant for lines as small as 4 μm, with slight increases beginning at 2 μm-long lines. The calculated sheet resistance then increases greatly for lines from 2 μm-200 nm. Below 200 nm, the calculated sheet resistance begins to drop.

The large-area sheet resistance was used to calculate expected resistances for the smaller

Iso length / μm	Mean measured R / kΩ	Calculated R_{sh} / (Ω/sq)	Expected R / kΩ
0.06	490.40	98.29	591.47
0.12	463.23	185.73	295.73
0.2	384.14	255.93	177.44
0.5	110.14	180.14	70.98
1	41.09	136.83	35.49
2	18.29	122.00	17.74
4	8.73	118.09	8.87
8	4.42	118.51	4.44
16	2.21	118.29	2.22

Table 9.4: Measured and calculated resistances for implant-isolated lines of various geometries. All lines were 300 μm wide.

lines of the geometries fabricated, and are plotted along with the measured resistances of the lines in Figure 9.16. It is clear that whilst the measurements are well-behaved above 1 μm, the resistances diverge as expected below this length. The peak difference between expected and actual resistances occurs at 200 nm, where the real resistances are over twice those expected. The resistances again converge below this, with the resistance of a 60 nm line, the minimum fabricated, actually around 100 kΩ smaller than expected.

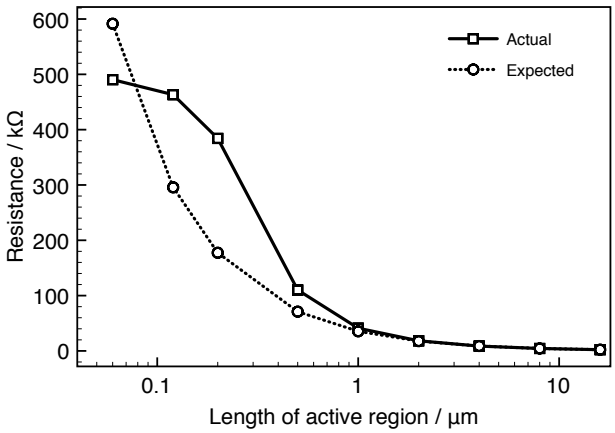


Figure 9.16: Expected and measured resistances for implant-isolated 300 μm-wide lines of varying length using the double krypton process.

The results are explicable by the simulated data. Recoil cascades extend far beyond the ionic lateral straggle as a consequence of the interactions of the massive Kr^+ ions with the lattice. As a result, whilst the ion penetration might isolate directly only a very short distance under the implant mask, as shown in the simulations of Figure 9.15,

lattice interactions continue up to 100 nm under the mask. These interactions result in the redistribution of atoms from various layers and the unintentional formation of compositionally-randomised quaternaries of random doping in the cascade regions.

As the implant mask shrinks laterally, the recoil regions account for a larger percentage of the masked region, and, due to the unchanged physics of the implant process, do not scale. For lines of decreasing length, as a result, the central portion will be unaffected by the implant as previously, but will be rapidly shrinking in length with respect to the length of the total masked region. Since the material isolated is designed to be low-resistance for devices, in general, structural changes result in increased resistivity. One would therefore expect there to be a low-resistance stripe in the centre of the line, bordered on both sides by relatively high-resistance quaternaries formed by the implant.

The following drop in resistance for shorter lengths is more difficult to explain given the argument just presented. As the length decreases below 200 nm, the entire region below the mask becomes subject to unintentional interactions as a result of implant-related recoil cascades as per the simulations of Figure 9.15. It is clear, however, that the bulk of the atomic displacements happen over a range of around 50 nm, with only the extreme cases resulting in 100 nm dislocations. As a result, for high-fluence implants such as that carried out, it is probable that for shorter lines such as the 60 nm case, the entire area might be affected, leaving a drastically altered layer structure from that designed. Since dopant moves as a result of the cascades, it is possible that the relative doping of the cap and channel changes such that the channel resistivity is drastically increased, leaving the bulk of the conduction only through the low-resistance cap and potentially creating a drop in resistivity.

There is, however, one further explanation. It is important to realise that implantation results in dopant straggle three-dimensionally around the implant point, with cascades resulting in atomic displacements in random directions. In many cases, the recoiling atoms backscatter in opposite directions to the start of the cascade, such that atoms originating under the mask can recoil to a point outwith it, back into the "damaged" region. It is therefore possible that, since the whole area under the mask is subject to recoil effects, the length of the low-resistance region might be larger than the mask if the region subject to recoils comprises a sufficient volume of the isolated region. Hence, for very small areas, the resistance could drop.

It is impossible to determine the exact causes of these effects without detailed study of the lattice by TEM, Rutherford Backscatter or Ion Channelling methods. It is, however,

clear that lateral effects become increasingly important when isolating small areas. If implant isolation were ever to be used to fabricate digital devices of CMOS proportions, these phenomena would present significant challenges.

9.6 Summary

This chapter has outlined the motivation for planarisation of HEMT structures, and the theory behind implant isolation, which is suggested as a potential method for achieving planar, well-isolated devices.

This work investigates previously-reported heavy-ion implantation, shown to be effective in bulk InP and InGaAs, for its effectiveness in a HEMT layer stack, where both narrow and wide bandgap materials are involved. Both krypton implants and iron implants were investigated, the krypton implant relying on damage-based isolation mechanisms, the iron ostensibly utilising chemical isolation techniques by lattice substitution. Double implants were required for both to effectively isolate a complete device layer stack.

Double krypton implants were shown to produce improved isolation of 1-2 MΩ/sq using 150 / 15 keV implants. Various annealing steps were investigated for the iron implants, up to a maximum temperature of 650 °C. Isolation was not shown to significantly improve with annealing; although sheet electron concentration decreased, some lattice healing appeared to occur. In addition, both the epitaxy and contacts were shown to be unstable at annealing temperatures above 500 °C.

Transmission lines and devices were fabricated using the double krypton implant, showing results broadly similar to the isolation achieved using conventional techniques. In particular, functional devices with performance identical to conventionally-isolated devices were realised. These devices yielded drain currents, transconductances and cutoff frequencies on a par with the best previously-realised metamorphic devices of a similar gate length.

Issues relating to implant-related intermixing of epitaxial composition were also discussed and experimentally verified. It was shown, in particular, that the lateral straggle of atoms as a consequence of implant-generated recoil cascades plays a particular role in the composition and resistance of small isolated structures, with unpredictable effects at nanometric length scales. Whilst unlikely to be a problem for HEMTs, predominantly used in r.f. applications, this may present major problems in potential future digital

devices.

Considering this success of the implant scheme in the fabrication of high-performance planar HEMTs, it appears to be a natural technology to apply to the 10-20 nm devices realised in Chapters 7 and 8 in order to increase device yield and reduce gate parasitic capacitance.

10. Conclusions and Future Work

The continued growth in applications exploiting the millimetre and sub-millimetre frequency bands for imaging and communications has increased the need for improved high-frequency transistor performance. As discussed in the introduction, the HEMT is a key low noise, high-frequency circuit component, and requires careful design to improve performance as the device is scaled.

This work has outlined a new strategy for the realisation of devices with gate lengths as small as 10 nm using a two-step gate lithography process. Using high-resolution electron beam lithography and reactive ion etching, these processes have allowed the fabrication of 22 nm gates with a sample yield close to 100%.

Two separate processes have additionally been developed for the fabrication of 10 nm gates using this technique, either by careful control of the etching and development conditions, or by the use of plasma deposition and etching of silicon nitride to reduce the gate length from larger features in a self-aligned manner. These techniques can therefore allow the fabrication of high-yield T-gates with very short gate lengths using a variety of fabrication tools, environments and processes.

The background issues underlying the realisation of high-performance short gate length devices have been explored in some depth, with a particular focus on the effect of various surface treatments on the gate recess region. In particular, the importance of minimising RIE etch times and powers has been noted, and particularly the importance of the cleaning processes used in the stripping of native oxides from the sample surface, which have been shown to acutely impact the device transfer characteristics. As a consequence, a revised process flow, involving multiple low-power RIE steps and a departure from the use of aggressive acids such as hydrofluoric acid, was developed.

These new processes resulted in the fabrication of well-behaved 22 nm devices with peak

transconductances as high as 1 S/mm and output currents of over 1.2 A/mm, with cutoff frequencies of around 360 GHz. Additionally, 50 nm devices fabricated using the same processes on the same material showed improved d.c. transconductance of up to 1.6 S/mm. It was particularly noted that device transconductance dropped off markedly with reducing gate length, underscoring the need for optimisation of the vertical device architecture. Additionally, the need for minimisation of the source and drain parasitic resistances was highlighted, noting the unusually high resistances arising from the device material used. Simple re-simulation, where these access resistances were halved, yielded cutoff frequencies of over 450 GHz, leaving all other parameters unchanged. Dielectric etching after the completion of the devices was also explored in order to reduce parasitic capacitances arising from the presence of silicon nitride adjacent to the gate. Although excessive dry etching resulted in damage to the devices, capacitances were reduced by 30%. In conjunction with the reduction of resistances envisaged by redesign of the material, device simulations resulted in cutoff frequencies higher than 600 GHz with no additional changes.

The material system was shown to have significant impact on device performance, with a need for sheet resistance reduction in conjunction with greatly reduced gate-channel separations to allow the fabrication of well-scaled devices at 22 nm gate lengths and below; changes expected to yield increased transconductance as well as reduced parasitic contributions. As a consequence, material was designed using two different doping techniques for the anticipated realisation of wafer structures with gate-channel separations as small as 2-4 nm. One method involved the retention of multiple dopant planes, a strategy conventionally used in HEMT fabrication, which yields high electron density, whilst the other involved the migration to backside channel doping only, allowing minimal barrier thickness. Although these structures were shown by simulation to yield suitable architectures and electron distributions, only the most conservative was fabricated during the timescale of this project.

Given the requirement for reduced source and drain resistances, strategies were also developed for the realisation of devices with source-drain gaps as small as 100 nm, yielding a wide variety of laterally-scaled device architectures. Although source-drain gaps as small as 40 nm were also realised, these proved to be constrained by the dimensions of the T-gate stack and recess used in the process flow. Functional devices with source-drain separations as small as 100 nm were realised using these processes, though there were additional problems, likely arising from dry etch damage following recalibration of the etch chemistry.

Processes were also developed for the effective implant isolation of high-indium HEMT wafers, using multiple krypton implants to define high-resistivity regions of semiconductor by physical damage mechanisms. Iron implants were additionally used, and shown to have negligible chemical isolation effects at device-compatible processing temperatures. Sheet resistances of up to 2 MΩ/sq were realised using the krypton process, and implant-isolated devices with identical performance to conventionally-isolated devices fabricated concurrently were demonstrated. Additionally, these devices were comparable to the highest-performance devices reported at similar gate lengths on metamorphic substrates.

Much of the work envisioned at the start of the project has been achieved, since high-performance devices with gate lengths of 22 nm have been fabricated. There remains, however, a need for the continued optimisation of the various epitaxial structures and parasitic components already explored in some depth. In particular, there is a need to integrate the various processes developed more completely to fully realise the potential of the combined processes.

Further work on highly-scaled HEMTs should therefore follow these priorities. Firstly, the cause of material damage to later devices, suspected to be issues related to the recalibration of the RIE system, must be verified and eliminated. The 10 nm gate etch time also requires to be recalibrated to ensure correct pattern transfer with the revised etch chemistry. At this stage, highly laterally-scaled devices can be fabricated with a range of source-drain separations at a range of gate lengths down to 10 nm on the latest low sheet resistance c577 wafer realised. From these devices, any further need for materials or process optimisation can be identified, since a broad spectrum of parasitic component values should result from the wide range of possible gate lengths and source-drain separations. The need for potential final dielectric etching and the integration of implant processes can then be explored if necessary to reduce parasitic capacitance or improve yield. The epitaxial realisation of device-quality wafers of short gate-channel separations is then of the utmost importance, as noted in the degradation of transconductance observed during the fabrication of 22 nm devices, and much can be learned from the fabrication of various device gate lengths on a variety of scaled epitaxial structures. Material development should therefore result in improved transconductance on the short gate length architectures developed in the work of this thesis.

In summary, this work has yielded significant improvements to the fabrication technologies that enable the realisation of highly-scaled High Electron Mobility Transistors. It is expected that, once fully integrated, the processes developed in the course of this work

will form a complete platform for the realisation of highly-scaled devices with leading high-frequency performance and a better understanding of the electron transport phenomena governing nanoscale transistor operation.

Appendices

A. Device Process Flows

This appendix contains the complete process flows used in the fabrication of the devices developed in the course of this work. Since each of the process flows uses generic modules developed designed to be interchangeable, each unique process is listed only once. The reader is referred to the relevant section for each.

A.1 22 nm Devices - Large source-drain gaps

Lithography step	**Process step**	**Process**
Markers	Clean	2hrs acetone, IPA rinse
	Spin	4% 2010 5k 60s, 2.5% 2041 5k 60s
	Bake	1hr, 2hrs 180°C
	Expose	Global markers: Dose 220 μCcm^{-2}, 64nA beam, VRU40 — Penrose markers: Dose 250 μCcm^{-2}, 1nA beam, VRU4
	Develop	2.5:1 IPA:MIBK, 60s, 23°C
	O_2 Ash	40W 60s
	De-oxidise	4:1 H_2O:HCl, 30s, 30s H_2O rinse
	Metallise	15 Ti 70 Au
	Liftoff	2hrs acetone, IPA rinse
Recess	Clean	2hrs acetone, IPA rinse
	Spin	2.5% 2041 5k 60s
	Bake	3hrs 180°C
	Expose	1nA beam, dose 1800 μCcm^{-2}, VRU10, Res 0.5nm
	Develop	2:1 IPA:MIBK, 60s, 23°C, IPA rinse
	O_2 Ash	40W 60s

Continued on Next Page...

Lithography step	Process step	Process
	Etch	De-oxidise first, 15s succinic
	Strip	2hrs acetone, IPA rinse
	Metrology	Optical and S4700 inpection
Plasma Deposition	Silicon nitride	50nm silicon nitride, SiH_4/N_2=6.2/6sccm, coil=100W, platen=0W, 4mTorr, 35°C, rate 11.5nm/min
Mark Protect	Clean	2hrs acetone, IPA rinse
	Spin	12% 2010 5k 60s
	Bake	30mins 180°C
	Expose	Dose 305 μCcm^{-2}, 64nA beam, VRU40
	Develop	2:1 IPA:MIBK, 60s, 23°C, IPA rinse
	O_2 Ash	40W 60s
Gate 1	Spin	50% ZEP520A, 5k, 60s
	Bake	40mins 180°C
	Expose	Dose 1700 μCcm^{-2}, 1nA beam, VRU10, Res 0.5nm
	Develop	o-xylene 30s, 23°C, IPA rinse
	Dry Etch	4m40 SF_6/N_2=5/55sccm, 20W, 15mTorr, 30°C, rate 16nm/min
	De-oxidise	10:1 H_2O:NH_3OH 20s, 30s H_2O rinse
	Metallise	15Ti 15Pt 15Au
	Liftoff	2hrs Microposit 1165
	Metrology	S4700
Isolation	Clean	2hrs acetone, IPA rinse
	Spin	12% 2010 5k 60s
	Bake	3hrs 180°C
	Expose	Dose 305 μCcm^{-2}, 64nA beam, VRU40
	Develop	2:1 IPA:MIBK, 60s, 23°C
	O_2 Ash	40W 60s
	Dry Etch	4m SF_6/N_2
	Bake	15m 120°C
	Wet etch	De-oxidise first, 90s 1:1:100 ortho:H_2O2:H_2O, 15s succinic
	Metrology	AFM, S4700, TLM, iso squares

Continued on Next Page...

Lithography step	Process step	Process
Gate 2	Clean	2hrs acetone, IPA rinse
	Spin	8% 2010 5k 60s, 1:4 LOR 5k 60s, 58% UVIII 3k 60s
	Bake	2hrs 180°C, 15m 180°C, 60s hotplate 126.7°C
	Expose	Gate: Dose 100 μCcm^{-2}, 1nA beam, VRU10, CFA36, Feeds: Dose 500 μCcm^{-2}, 64nA beam, VRU40
	Postbake	90s hotplate 126.7°C
	Develop	60s CD26 room temperature, 90s o-xylene, 23°C
	O_2 Ash	40W 30s
	De-oxidise	4:1 H_2O:HCl, 30s, 30s H_2O rinse
	Metallise	15Ti 15Pt 180Au
	Liftoff	2hrs acetone, IPA rinse
	Clean	5m Microposit 1165, H_2O rinse
Ohmic	Clean	2hrs acetone, IPA rinse
	Spin	4% 2010 5k 60s, 2.5% 2041 5k 60s
	Bake	1hr, 2hrs 180°C
	Expose	Dose 305 μCcm^{-2}, 64nA beam, VRU40
	Develop	2:1 IPA:MIBK 60s, 23°C, IPA rinse
	O_2 Ash	40W, 60s
	Dry Etch	4m SF_6/N_2
	De-oxidise	4:1 H_2O:HCl, 30s, 30s H_2O rinse
	Metallise	11Au 11Ge 11Au 11Ge 20Au 12Ni 80Au
	Liftoff	2hrs acetone, IPA rinse
	Metrology	S4700, TLM, VdP
Bondpad	Clean	2hrs acetone, IPA rinse
	Spin	15% 2010 3k 60s, 4% 2041 5k 60s
	Bake	1hr/1hr 120°C
	Expose	Dose 305 μCcm^{-2}, 64nA beam, VRU40
	Develop	2:1 IPA:MIBK 60s, 23°C
	O_2 Ash	40W 60s
	Dry etch	4m SF_6/N_2
	De-oxidise	4:1 H_2O:HCl, 30s, 30s H_2O rinse
	Metallise	50 NiCr 1200 Au (50NiCr 400 Au short run)

Continued on Next Page...

Lithography step	Process step	Process
	Liftoff	2hrs acetone, IPA rinse
	Metrology	S4700, TLM, VdP, DC, RF, S900

A.2 22 nm Devices - Short source-drain gaps

Lithography step	**Process step**	**Process**
Markers	all steps	as for Section A.1
Recess	all steps	as for Section A.1
Plasma Deposition	all steps	as for Section A.1
Mark Protect	all steps	as for Section A.1
Gate 1	all steps	as for Section A.1
Isolation	all steps	as for Section A.1
Ohmic	Clean	2hrs acetone, IPA rinse
	Spin	4% 2010 5k 60s, 2.5% 2041 5k 60s
	Bake	1hr, 2hrs 180°C
	Expose	2μm ohmics: Dose 220 μCcm^{-2}, 64nA beam, VRU40 — Short gap ohmics: Dose 220 μCcm^{-2}, 8nA beam, VRU12
	Develop	2.5:1 IPA:MIBK 60s, 23°C, IPA rinse
	O_2 Ash	40W, 60s
	Dry Etch	4m SF_6/N_2
	De-oxidise	4:1 H_2O:HCl, 30s, 30s H_2O rinse
	Metallise	9Au 9Ge 9Au 8Ni 15Au
	Liftoff	2hrs acetone, IPA rinse
	Metrology	S4700, TLM, VdP
Gate 2	all steps	as for Section A.1
Bondpad	all steps	as for Section A.1

A.3 10 nm Devices

Lithography step	Process step	Process
Markers	all steps	as for Section A.1 or A.2
Recess	all steps	as for Section A.1 or A.2
Plasma Deposition	Silicon nitride	20nm silicon nitride, SiH_4/N_2=6.2/6sccm, coil=100W, platen=0W, 4mTorr, 35°C, rate 11.5nm/min
Mark Protect	all steps	as for Section A.1 or A.2
Gate 1	Spin	50% ZEP520A, 5k, 60s
	Bake	40mins 180°C
	Expose	Dose 600 μCcm^{-2}, 1nA beam, VRU10, Res 0.5nm
	Develop	o-xylene 30s, 23°C, IPA rinse
	Dry Etch	105s SF_6/N_2=5/55sccm, 20W, 15mTorr, 30°C, rate 16nm/min
	De-oxidise	10:1 H_2O:NH_3OH 20s, 30s H_2O rinse
	Metallise	15Ti 5Pt 5Au
	Liftoff	2hrs Microposit 1165, H_2O rinse
	Metrology	S4700
Isolation	all steps	as for Section A.1 or A.2
Ohmic	all steps	as for Section A.1 or A.2
Gate 2	all steps	as for Section A.1 or A.2
Bondpad	all steps	as for Section A.1 or A.2

A.4 Implanted Devices

Lithography step	Process step	Process
Markers	all steps	as for Section A.1
Plasma Deposition	Silicon nitride	400nm silicon nitride, SiH_4/N_2=6.2/6sccm, coil=100W, platen=0W, 4mTorr, 35°C, rate 11.5nm/min
Implant mask	Spin	100% NEB-31, 3k, 60s
	Bake	2mins hotplate 87°C
	Expose	Dose 30 μCcm^{-2}, 64nA beam, VRU40, Res 1.25nm
	Postbake	2mins hotplate 97°C
	Develop	CD26, room temperature, H_2O rinse
	Dry Etch	SF_6/N_2=5/55sccm, 20W, 15mTorr, 30°C, rate 16nm/min, reflectometry-terminated
	Metrology	Optical, AFM
Implant	Implant	Various conditions
	Mask etch	SF_6/N_2=5/55sccm, 20W, 15mTorr, 30°C, rate 16nm/min, reflectometry-terminated
	Metrology	Optical Nomarski, AFM
Gate 2	Clean	2hrs acetone, IPA rinse
	Spin	2.5% 2041 3.7k 60s, 1:4 LOR 5k 60s, 58% UVIII 3.5k 60s
	Bake	2hrs 180°C, 15m 180°C, 60s hotplate 126.7°C
	Expose	Gate: Dose 100 μCcm^{-2}, 1nA beam, VRU10, CFA15, Feeds: Dose 500 μCcm^{-2}, 64nA beam, VRU40
	Postbake	90s hotplate 126.7°C
	Develop	60s CD26 room temperature, 60s o-xylene 23°C
	O_2 Ash	40W 30s
	De-oxidise	4:1 H_2O:HCl, 30s, 30s H_2O rinse
	Metallise	15Ti 15Pt 180Au
	Liftoff	2hrs acetone, IPA rinse
	Clean	5m Microposit 1165, H_2O rinse
Ohmic	all steps	as for Section A.1

Continued on Next Page...

Lithography step	**Process step**	**Process**
Bondpad	all steps	as for Section A.1

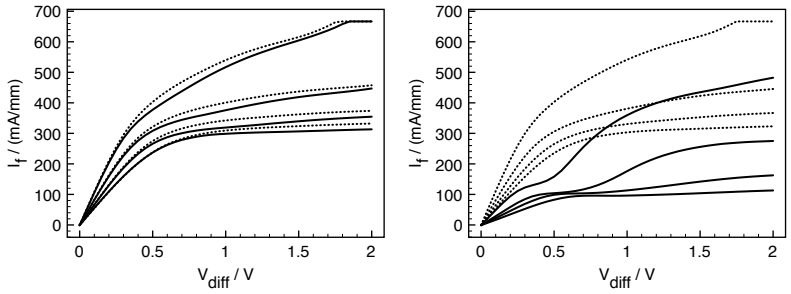
B. InAlAs surface treatments: Complete I-V results

This appendix provides more complete results to the InAlAs processing experiments summarised in Section 7.9. These measurements explored firstly a range of SF_6, HF and sulphide treatments, detailed in Figure B.1.

These results suggested the significant influence of HF, and so further experiments were carried out, shown in Figure B.2.

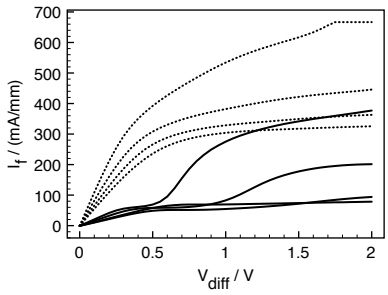
A range of acids and bases were then tested, shown in Figure B.3.

Each set of experiments used a set of standard TLM structures, with varying gap lengths from 2.5 - 5.5 μm.

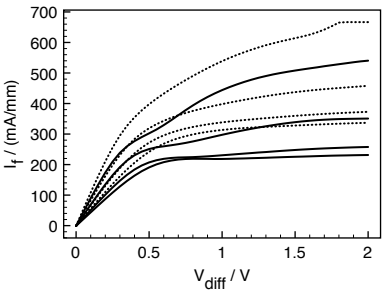


(a) SF_6 only

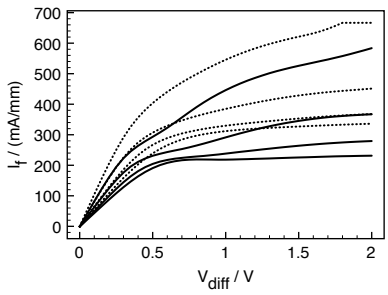
(b) 100:1 HF only



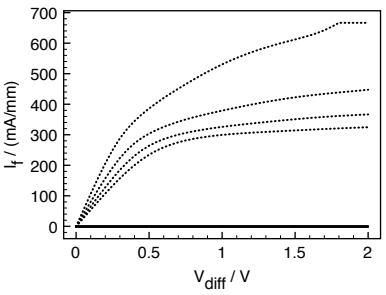
(c) 100:1 HF & SF_6



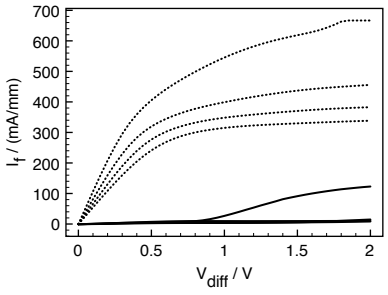
(d) Sulphidation & SF_6



(e) Sulphidation & bake & SF_6

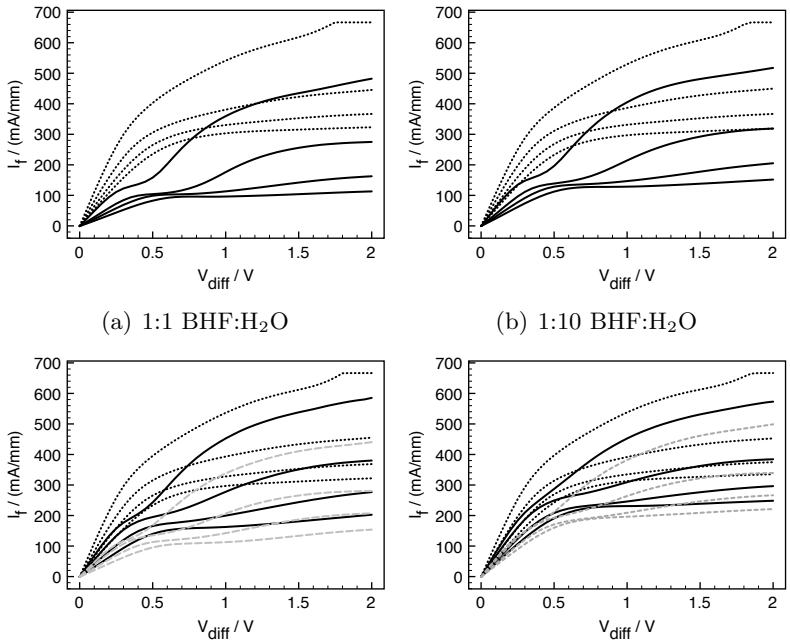


(f) 100:1 HF & Sulphidation & SF_6



(g) 100:1 HF & Sulphidation & bake & SF_6

Figure B.1: Comparison of various recessed TLM sites before and after various post-recessing surface treatments. Gap size increases from top trace to bottom from 2.5 μm to 5.5 μm in 1 μm steps. Dashed traces are the pre-treatment curves in each case.



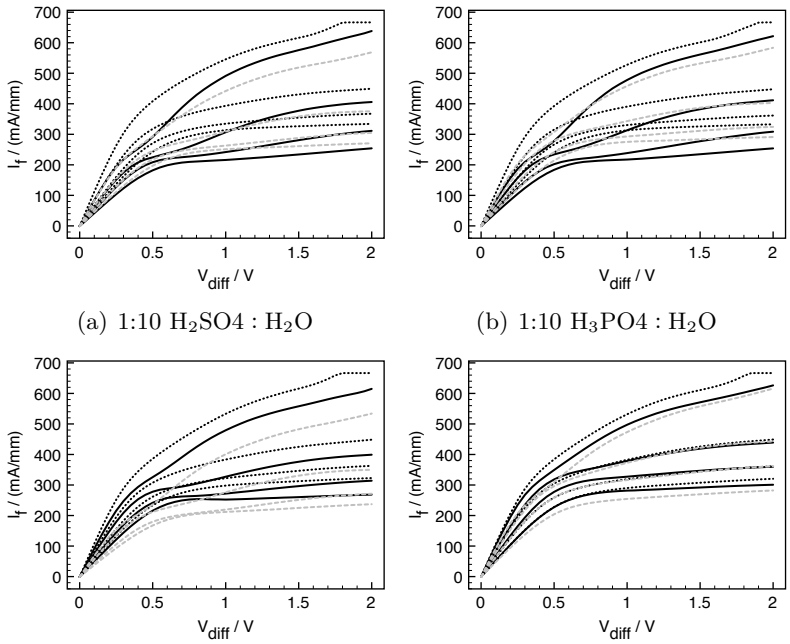
(a) 1:1 BHF:H_2O

(b) 1:10 BHF:H_2O

(c) 1:100 BHF:H_2O

(d) 1:100 HF:H_2O

Figure B.2: Comparison of various recessed TLM sites before and after various HF-based surface treatments. Dashed traces are the pre-treatment curves in each case.



(a) 1:10 $H_2SO4 : H_2O$

(b) 1:10 $H_3PO4 : H_2O$

(c) 1:4 $HCl : H_2O$

(d) 1:10 $NH_3OH : H_2O$

Figure B.3: Comparison of various recessed TLM sites before and after various non-HF surface treatments.

References

[1] T. Mimura, "The early history of the high electron mobility transistor (HEMT)," *IEEE Transactions on Microwave Theory and Techniques*, vol. 50, pp. 780–782, **2002**.

[2] R. Dingle, H. Stormer, A. Gossard and W. Wiegmann, "Electron mobilities in modulation-doped semiconductor heterojunction superlattices," *Applied Physics Letters*, vol. 33, pp. 665–667, **1978**.

[3] M. Shur, "Terahertz technology: devices and applications," in *Solid-State Device Research Conference*, pp. 13–21 (**2005**).

[4] M. Rodwell, E. Lind, Z. Griffith, S. Bank, A. Crook, U. Singisetti, M. Wistey, G. Burek and A. Gossard, "Frequency Limits of InP-based Integrated Circuits," in *International Conference on Indium Phosphide and Related Materials*, pp. 9–13 (**2007**).

[5] W. Hafez and M. Feng, "Experimental demonstration of pseudomorphic heterojunction bipolar transistors with cutoff frequencies above 600 GHz," *Applied Physics Letters*, vol. 86, p. 152 101, **2005**.

[6] M. Feng and W. Snodgrass, "InP Pseudomorphic Heterojunction Bipolar Transistor (PHBT) With $f_t >$ 750GHz," in *International Conference on Indium Phosphide and Related Materials*, pp. 399–402 (**2007**).

[7] D. Pavlidis, "HBT vs. PHEMT vs. MESFET: What's best and why," in *Digest of the International Conference on Compound Semiconductor Manufacturing Technology* (**1999**).

[8] D. Streit, R. Lai, A. Oki and A. Gutierrez-Aitken, "InP HEMT and HBT applications beyond 200 GHz," in *International Conference on Indium Phosphide and Related Materials* (**2002**).

[9] J. Lynch, J. Schulman, J. Schaffner, H. Moyer, Y. Royter, P. Macdonald and B. Hughes, "Low noise radiometers for passive millimeter wave imaging," in *International Conference on Infrared, Millimeter and Terahertz Waves*, pp. 1 – 3 (**2008**).

[10] B. Kim, K. Lee and D. Yu, "Current Status of Millimeter-Wave Transistor Technology," in *Topical Symposium on Millimeter Waves* (**2004**).

[11] C. Kuo, H. Hsu, E. Chang, C. Chang and Y. Miyamoto, "RF and Logic Performance Improvement of $In_{0.7}Ga_{0.3}As/InAs/In_{0.7}Ga_{0.3}As$ Composite Channel HEMT Using Gate Sinking Technology," *IEEE Electron Device Letters*, vol. 29, pp. 290–293, **2008**.

[12] D.-H. Kim and J. del Alamo, "30-nm InAs Pseudomorphic HEMTs on an InP Substrate With a Current-Gain Cutoff Frequency of 628 GHz," *Electron Device Letters*, vol. 29, pp. 830–833, **2008**.

[13] N. Waldron, D.-H. Kim and J. del Alamo, "90 nm Self-aligned Enhancement-mode InGaAs HEMT for Logic Applications," in *IEEE International Electron Devices Meeting*, pp. 633–636 (**2007**).

[14] M. Passlack, K. Rajagopalan, J. Abrokwah and R. Droopad, "Implant-free high-mobility flatband MOSFET: principles of operation," *IEEE Transactions on Electron Devices*, vol. 53, pp. 2454–2459, **2006**.

[15] R. Hill, D. Moran, X. Li, H. Zhou, D. Macintyre, S. Thoms, A. Asenov, P. Zurcher, K. Rajagopalan, J. Abrokwah, R. Droopad, M. Passlack and I. Thayne, "Enhancement-Mode GaAs MOSFETs With an $In_{0.3}Ga_{0.7}As$ Channel, a Mobility of over 5000 cm^2/Vs and Transconductance of over 475 $\mu S/\mu m$," *Electron Device Letters*, vol. 28, pp. 1080–1082, **2007**.

[16] R. Hill, R. Droopad, D. Moran, X. Li, H. Zhou, D. Macintyre, S. Thoms, O. Ignatova, A. Asenov, K. Rajagopalan, P. Fejes, I. Thayne and M. Passlack, "1 μm gate length, $In_{0.75}Ga_{0.25}As$ channel, thin body n-MOSFET on InP substrate with transconductance of 737 $\mu S/\mu m$." *Electronics Letters*, vol. 44, pp. 498–500, **2008**.

[17] Y. Sun, E. Kiewra, J. de Souza, J. Bucchignano, K. Fogel, D. Sadana and G. Shahidi, "High-Performance $In_{0.7}Ga_{0.3}As$-Channel MOSFETs With κ Gate Dielectrics and α-Si Passivation," *Electron Device Letters*, vol. 30, pp. 5–7, **2009**.

[18] M. Lundstrom, *Fundamentals of Carrier Transport*, pp. 13–20 (Addison-Wesley, Wokingham, **1990**).

[19] M. Shur, *Introduction to Electronic Devices*, pp. 6–8 (John Wiley & Sons, Chichester, **1996**).

[20] —, *Introduction to Electronic Devices*, pp. 72–74 (John Wiley & Sons, Chichester, **1996**).

[21] S. Sze, *Physics of Semiconductor Devices* (John Wiley & Sons, New York, **1981**).

[22] J. Ayubi-Moak, D. Ferry, S. Goodnick and R. Akis, "Simulation of Ultrasubmicrometer-Gate $In_{0.52}Al_{0.48}As/In_{0.75}Ga_{0.25}As/In_{0.52}Al_{0.48}As/InP$ Pseudomorphic HEMTs Using a Full-Band Monte Carlo Simulator," *IEEE Transactions on Electron Devices*, vol. 54, pp. 2327–2338, **2007**.

[23] M. Shur, *Introduction to Electronic Devices*, pp. 26–27 (John Wiley & Sons, Chichester, **1996**).

[24] —, *Introduction to Electronic Devices*, p. 75 (John Wiley & Sons, Chichester, **1996**).

[25] S. Tiwari, *Compound Semiconductor Device Physics* (Academic Press, London, **1992**).

[26] M. Lundstrom, *Fundamentals of Carrier Transport* (Cambridge University Press, Cambridge, **2000**).

[27] F. Capotondi, G. Biasiol, D. Ercolani and L. Sorba, "Scattering mechanisms in undoped InGaAs/InAlAs two-dimensional electron gases," *Journal of Crystal Growth*, vol. 278, pp. 538–543, **2005**.

[28] B. Nag and M. Das, "Scattering potential for interface roughness scattering," *Applied Surface Science*, vol. 182, pp. 357–360, **2001**.

[29] M. Shur, *Introduction to Electronic Devices*, pp. 137–139 (John Wiley & Sons, Chichester, **1996**).

[30] R. Dorf, *The Electrical Engineering Handbook*, p. 991 (CRC Press, London, **1997**).

[31] D. Ferry, J. Barker and H. Grubin, "Hot-carrier constraints on transient transport in very small semiconductor devices," *IEEE Transactions on Electron Devices*, vol. 28, pp. 905–911, **1981**.

[32] S. Tiwari, *Compound Semiconductor Device Physics*, pp. 68–69 (Academic Press, London, **1992**).

[33] A. Katz, *Indium Phosphide and Related Materials : Processing, Technology and Devices* (Artech House, Boston, **1992**).

[34] E. Kobayashi, C. Hamaguchi, T. Matsuoka and K. Taniguchi, "Monte Carlo study of hot-electron transport in an InGaAs/InAlAs single heterostructure," *IEEE Transactions on Electron Devices*, vol. 36, pp. 2353–2360, **1989**.

[35] A. Khalid, N. Pilgrim, G. Dunn, M. Holland, C. Stanley, I. Thayne and D. Cumming, "A Planar Gunn Diode Operating Above 100 GHz," *Electron Device Letters*, vol. 28, pp. 849 – 851, **2007**.

[36] I. Fritz, S. Picraux, L. Dawson and T. Drummond, "Dependence of critical layer thickness on strain for InGaAs/GaAs strainedlayer superlattices," *Applied Physics Letters*, vol. 46, pp. 967–969, **1985**.

[37] A. Denton and N. Ashcroft, "Vegard's law," *Physical Review A*, vol. 43, pp. 3161–3164, **1991**.

[38] I. Vurgaftman, J. Meyer and L. Ram-Mohan, "Band parameters for III–V compound semiconductors and their alloys," *Journal of Applied Physics*, vol. 89, pp. 5815–5875, **2001**.

[39] C. Köpf, H. Kosina and S. Selberherr, "Physical models for strained and relaxed GaInAs alloys: Band structure and low-field transport," *Solid State Electronics*, vol. 41, pp. 1139–1152, **1997**.

[40] S. Babiker, A. Asenov, S. Roy and S. Beaumont, "Strain engineered pHEMTs on virtual substrates: a Monte Carlo simulation study," *Solid-State Electronics*, vol. 43, p. 1281, **1999**.

[41] R. Anderson, "Experiments on Ge-GaAs heterojunctions," *Solid-State Electronics*, vol. 5, pp. 341–344, **1962**.

[42] W. R. Frensley and H. Kroemer, "Theory of the energy-band lineup at an abrupt semiconductor heterojunction," *Physical Review B*, vol. 16, pp. 2642–2652, **1977**.

[43] J. Shay, S. Wagner and J. Phillips, "Heterojunction band discontinuities," *Applied Physics Letters*, vol. 28, pp. 31–33, **1976**.

[44] J. Tersoff, "Schottky Barrier Heights and the Continuum of Gap States," *Physical Review Letters*, vol. 52, p. 465, **1984**.

[45] J. Davies, *The Physics of Low-Dimensional Semiconductors* (Cambridge University Press, Cambridge, **1998**).

[46] I. Tan, G. Snider, L. Chang and E. Hu, "A selfconsistent solution of Schrödinger–Poisson equations using a nonuniform mesh," *Journal of Applied Physics*, vol. 68, pp. 4071–4076, **1990**.

[47] Y. Ando and T. Itoh, "Analysis of charge control in pseudomorphic two-dimensional electron gas field-effect transistors," *Electron Devices*, vol. 35, pp. 2295–2301, **1988**.

[48] I. Tamm, "A possible kind of electron binding on crystal surfaces," *Physikalische Zeitschrift der Sowjetunion*, vol. 1, p. 733, **1932**.

[49] W. Shockley, "On the Surface States Associated with a Periodic Potential," *Physical Review*, vol. 56, pp. 317–323, **1939**.

[50] J. Bardeen, "Surface states and rectification at a metal semi-conductor contact," *Physical Review*, vol. 71, pp. 717–727, **1947**.

[51] V. Heine, "Theory of Surface States," *Physical Review*, vol. 138, pp. 1689–1696, **1965**.

[52] W. Schottky, "Vereinfachte und erweiterte Theorie der Randschichtgleichrichter," *Zeitschrift für Physik A: Hadrons und Nuclei*, pp. 539–592, **1942**.

[53] A. Cowley and S. Sze, "Surface States and Barrier Height of Metal-Semiconductor Systems," *Journal of Applied Physics*, vol. 36, pp. 3212–3220, **1965**.

[54] E. Skuras and C. Stanley, "Fermi energy pinning at the surface of high mobility $In_{0.53}Ga_{0.47}AsIn_{0.52}Al_{0.48}As$ modulation doped field effect transistor structures," *Applied Physics Letters*, vol. 90, pp. 133 506–1 – 3, **2007**.

[55] E. Skuras, G. Pennelli, A. Long and C. Stanley, "Molecular-beam epitaxy growth of InGaAs–InAlAs high electron mobility transistors with enhanced electron densities and measurement of InAlAs surface potential," *Journal of Vacuum Science and Technology B: Microelectronics and Nanometer Structures*, vol. 19, pp. 1524–1528, **2001**.

[56] S. Tiwari, *Compound Semiconductor Device Physics*, p. 216 (Academic Press, London, **1992**).

[57] C. Peng, T. Won, C. Litton and H. Morkoc, "A high-performance InGaAs/InAlAs double-heterojunction bipolar transistor with nonalloyed n^+-InAs cap layer on InP(n) Grown by Molecular Beam Epitaxy," *Electron Device Letters*, vol. 9, pp. 331–333, **1988**.

[58] D. Moran, H. McLelland, K. Elgaid, G. Whyte, C. Stanley and I. Thayne, "50-nm self-aligned and "standard" T-gate InP pHEMT comparison: The influence of parasitics on performance at the 50-nm node," *IEEE Transactions on Electron Devices*, vol. 53, pp. 2920–2925, **2006**.

[59] A. Wakita, N. Moll, A. Fischer-Colbrie and W. Stickle, "Design and surface chemistry of nonalloyed ohmic contacts to pseudomorphic InGaAs on n^+GaAs," *Journal of Applied Physics*, vol. 68, pp. 2833–2838, **1990**.

[60] N. Braslau, "Alloyed ohmic contacts to GaAs," *Journal of Vacuum Science and Technology B: Microelectronics and Nanometer Structures*, vol. 19, p. 803, **1981**.

[61] A. Baca, "A survey of ohmic contacts to III-V compound semiconductors," *Thin Solid Films*, vol. 308-309, pp. 599–606, **1997**.

[62] W. Patrick, W. Mackie, S. Beaumont and C. Wilkinson, "Low-temperature annealed contacts to very thin GaAs epilayers," *Applied Physics Letters*, vol. 48, pp. 986–987, **1986**.

[63] T. Shen, G. Gao and H. Morkoç, "Recent developments in ohmic contacts for III–V compound semiconductors," *Journal of Vacuum Science and Technology B: Microelectronics and Nanometer Structures*, vol. 10, pp. 2113–2132, **1992**.

[64] D. Delagebeaudeuf and N. Linh, "Metal-(n) AlGaAs-GaAs two-dimensional electron gas FET," *Electron Devices*, vol. 29, pp. 955–960, **1982**.

[65] L. Nguyen, L. Larson and U. Mishra, "Ultra-high speed modulation-doped field-effect transistors: a tutorial review," in *Proceedings of the IEEE*, vol. 80, pp. 494–518 (**1992**).

[66] P. Ladbrooke, *MMIC Design : GaAs FETs and HEMTs*, pp. 125–135 (Artech House, Boston, **1989**).

[67] S. Bahl, J. D. Alamo and C. MIT, "Physics of breakdown in InAlAs/n-InGaAs heterostructurefield-effect transistors," *Electron Devices*, vol. 41, pp. 2268–2275, **1994**.

[68] J. Dickmann, S. Schildberg, K. Riepe and B. Maile, "Breakdown Mechanisms in Pseudomorphic InAlAs/$In_xGa_{1-x}As$ High Electron Mobility Transistors on InP. I:Off-State," *Japanese Journal of Applied Physics*, vol. 34, pp. 66–71, **1995**.

[69] J. Dickmann, S. Schildberg, A. Geyer and B. Maile, "Breakdown mechanisms in the on-state mode of operation of InAlAs/$In_xGa_{1-x}As$ pseudomorphic HEMTs," in *International Conference on Indium Phosphide and Related Materials*, pp. 335–338 (**1994**).

[70] J. Dickmann, S. Schildberg, K. Riepe and B. Maile, "Breakdown Mechanisms in Pseudomorphic InAlAs/$In_xGa_{1-x}As$ High Electron Mobility Transistors on InP. II:On-State," *Japanese Journal of Applied Physics*, vol. 34, pp. 1805–1808, **1995**.

[71] M. Somerville, R. Blanchard, J. D. Alamo and G. Duh, "A new gate current extraction technique for measurement of on-state breakdown voltage in HEMTs," *Electron Device Letters*, vol. 19, pp. 405–407, **1998**.

[72] G. Meneghesso, A. Neviani and R. Oesterholt, "On-state and off-state breakdown in GaInAs/InP composite-channel HEMTs with variable GaInAs channel thickness," *Electron Devices*, vol. 46, pp. 2–9, **1999**.

[73] A. Hulsmann, W. Bronner, K. Kohler and M. Baeumler, "The problem of breakdown in MODFETs," *Integrated Nonlinear Microwave and Millimeterwave Circuits*, pp. 149–152, **1994**.

[74] M. Somerville, J. del Alamo, W. Hoke and C. MIT, "Direct correlation between impact ionization and the kink effect inInAlAs/InGaAs HEMTs," *Electron Device Letters*, vol. 17, pp. 473–475, **1996**.

[75] M. Somerville, A. Ernst and J. del Alamo, "A physical model for the kink effect in InAlAs/InGaAs HEMTs," *IEEE Transactions on Electron Devices*, vol. 47, p. 922, **2000**.

[76] T. Suemitsu, T. Enoki, N. Sano, M. Tomizawa and Y. Ishii, "An analysis of the kink phenomena in InAlAs/InGaAs HEMT's using two-dimensional device simulation," *IEEE Transactions on Electron Devices*, vol. 45, pp. 2390–2399, **1998**.

[77] G. Zhou, A. Fischer-Colbrie, J. Miller and Y. Pao, "High output conductance of InAlAs/InGaAs/InP MODFET due to weak impact ionization in the InGaAs Channel," in *IEEE International Electron Devices Meeting*, pp. 247–250 (**1991**).

[78] W. Kruppa and J. Boos, "Low-frequency transconductance dispersion in InAlAs/InGaAs/InP HEMT's with single- and double-recessed gate structures," *IEEE Transactions on Electron Devices*, vol. 44, pp. 687–692, **1997**.

[79] T. Suemitsu, H. Yokoyama, Y. Ishii, T. Enoki, G. Meneghesso and E. Zanoni, "30-nm Two-Step Recess Gate InP-Based InAlAs/InGaAs HEMTs," *IEEE Transactions on Electron Devices*, vol. 49, pp. 1694–1700, **2002**.

[80] P. Ladbrooke, *MMIC Design : GaAs FETs and HEMTs*, pp. 91–124 (Artech House, Boston, **1989**).

[81] P. Wolf, "Microwave Properties of Schottky-barrier Field-effect Transistors," *IBM Journal of Research and Development*, vol. 14, pp. 125–141, **1970**.

[82] M. Shur, *Introduction to Electronic Devices*, pp. 391–393 (John Wiley & Sons, Chichester, **1996**).

[83] P. Ladbrooke, *MMIC Design : GaAs FETs and HEMTs*, pp. 206–207 (Artech House, Boston, **1989**).

[84] M. Shur, *Introduction to Electronic Devices*, p. 400 (John Wiley & Sons, Chichester, **1996**).

[85] P. Tasker and B. Hughes, "Importance of source and drain resistance to the maximum f_t of millimeter-wave MODFETs," *Electron Device Letters*, vol. 10, p. 291, **1989**.

[86] N. Moll, M. Hueschen and A. Fischer-Colbrie, "Pulse-doped AlGaAs/InGaAs pseudomorphic MODFETs," *IEEE Transactions on Electron Devices*, vol. 35, pp. 879 – 886, **1988**.

[87] S. Yeon, M. Park, J. Choi and K. Seo, "610 GHz InAlAs/$In_{0.75}$GaAs Metamorphic HEMTs with an Ultra-Short 15-nm-Gate," in *Electron Devices Meeting*, pp. 613–616 (**2007**).

[88] P. Ladbrooke, *MMIC Design : GaAs FETs and HEMTs*, pp. 223–224 (Artech House, Boston, **1989**).

[89] M. Das, "A high aspect ratio design approach to millimeter-wave HEMT structures," *IEEE Transactions on Electron Devices*, vol. 32, pp. 11–17, **1985**.

[90] K. Kalna and A. Asenov, "Tunnelling and Impact Ionization in Scaled Double Doped PHEMTs," in *European Solid-State Device Research Conference*, p. 303 (**2002**).

[91] S. Teitel and J. Wilkins, "Ballistic transport and velocity overshoot in semiconductors: Part I—Uniform field effects," *IEEE Transactions on Electron Devices*, vol. 30, pp. 150–153, **1983**.

[92] I. Kizilyalli, K. Hess, J. Larson and D. Widiger, "Scaling properties of high electron mobility transistors," *IEEE Transactions on Electron Devices*, vol. 33, p. 1427, **1986**.

[93] J. Han and D. Ferry, "Scaling of gate length in ultra-short channel heterostructure field effect transistors," *Solid-State Electronics*, vol. 43, p. 335, **1999**.

[94] M. Shur and L. Eastman, "Ballistic transport in semiconductor at low temperatures for low-power high-speed logic," *IEEE Transactions on Electron Devices*, vol. 26, pp. 1677–1683, **1979**.

[95] L. Eastman, R. Stall, D. Woodard, N. Dandekar, C. Wood, M. Shur and K. Board, "Ballistic electron motion in GaAs at room temperature," *Electronics Letters*, vol. 16, pp. 524 – 525, **1980**.

[96] M. Lundstrom and Z. Ren, "Essential physics of carrier transport in nanoscale MOSFETs," *IEEE Transactions on Electron Devices*, vol. 49, pp. 133 – 141, **2002**.

[97] A. Rahman, G. Klimeck and M. Lundstrom, "Novel channel materials for ballistic nanoscale MOSFETs-bandstructure effects," in *IEEE International Electron Devices Meeting*, p. 4 (**2005**).

[98] P. Solomon, S. Laux, I. Center and Y. Heights, "The ballistic FET: design, capacitance and speed limit," in *Electron Devices Meeting* (**2001**).

[99] J. Wang and M. Lundstrom, "Ballistic transport in high electron mobility transistors," *IEEE Transactions on Electron Devices*, vol. 50, pp. 1604–1609, **2003**.

[100] T. Maloney, "Polar mode scattering in ballistic transport GaAs devices," *Electron Device Letters*, vol. 1, pp. 54 – 54, **1980**.

[101] M. Shur and L. Eastman, "Ballistic and near ballistic transport in GaAs," *Electron Device Letters*, vol. 1, pp. 147–148, **1980**.

[102] —, "Near ballistic electron transport in GaAs devices at 77 K," *Solid-State Electron*, vol. 24, pp. 11–18, **1981**.

[103] K. Kalna, S. Roy, A. Asenov, K. Elgaid and I. Thayne, "Scaling of pseudomorphic high electron mobility transistors to decanano dimensions," *Solid-State Electronics*, vol. 46, p. 631, **2002**.

[104] S. Kang, "Ballistic transport at GHz frequencies in ungated HEMT structures," *Solid-State Electronics*, vol. 48, pp. 2013–2017, **2004**.

[105] J. Hauser, "Characteristics of junction field effect devices with small channel length-to-width ratios," *Solid-State Electronics*, vol. 10, p. 577, **1967**.

[106] T. Suemitsu, T. Enoki, H. Yokoyama and Y. Ishii, "Improved recessed-gate structure for sub-0.1-μm-gate InP-based High Electron Mobility Transistors," *Japanese Journal of Applied Physics*, vol. 37, pp. 1365–1372, **1998**.

[107] R. Akis, J. Ayubi-Moak, N. Faralli and D. Ferry, "The Upper Limit of the Cutoff Frequency in Ultrashort Gate-Length InGaAs/InAlAs HEMTs: A New Definition of Effective Gate Length," *Electron Device Letters*, vol. 29, pp. 306–308, **2008**.

[108] D. K. Ferry, J. Ayubi-Moak, R. Akis, N. Faralli, M. Saraniti and S. M. Goodnick, "Full-band CMC simulations of terahertz HEMTs," in *Journal of Physics: Conference Series*, vol. 109, pp. 012 001–1 – 6 (**2008**).

[109] Y. Awano, M. Kosugi, K. Kosemura, T. Mimura and M. Abe, "Short-channel effects in subquarter-micrometer-gate HEMTs: simulation and experiment," *IEEE Transactions on Electron Devices*, vol. 36, p. 2260, **1989**.

[110] M. Lundstrom, "Device physics at the scaling limit: what matters?" in *IEEE International Electron Devices Meeting*, pp. 33.1.1 – 33.1.4 (**2003**).

[111] H. Wong, "Beyond the conventional transistor," *IBM Journal of Research and Development*, vol. 46, pp. 133–168, **2002**.

[112] N. Wichmann, I. Duszynski, S. Bollaert, X. Wallart and A. Cappy, "InAlAs/InGaAs double-gate HEMTs with high extrinsic transconductance," in *International Conference on Indium Phosphide and Related Materials*, p. 295 (**2004**).

[113] N. Wichmann, I. Duszynski, X. Wallart, S. Bollaert and A. Cappy, "Fabrication and characterization of 100-nm $In_{0.53}Ga_{0.47}As - In_{0.52}Al_{0.48}As$ double-gate HEMTs with two separate gate controls," *Electron Device Letters*, vol. 26, p. 601, **2005**.

[114] T. Bryllert, L. Wernersson, L. Froberg and L. Samuelson, "Vertical high-mobility wrap-gated InAs nanowire transistor," *Electron Device Letters*, vol. 27, pp. 323–325, **2006**.

[115] A. Forchel, M. Scheffler, W. Riess and B. Ohlsson, "Nanowire-based one-dimensional electronics," *Materials Today*, vol. 9, pp. 28–35, **2006**.

[116] V. Schmidt, H. Riel, S. Senz, S. Karg and W. Riess, "Realization of a Silicon Nanowire Vertical Surround-Gate Field-Effect Transistor," *Small*, vol. 2, pp. 85–88, **2006**.

[117] R. Wang, J. Zhuge, R. Huang, Y. Tian, H. Xiao and L. Zhang, "Analog/RF Performance of Si Nanowire MOSFETs and the Impact of Process Variation," *Electron Devices*, vol. 54, pp. 1288–1294, **2007**.

[118] R. Williams, *Modern GaAs processing methods*, pp. 32–39 (Artech House, Boston, **1990**).

[119] —, *Modern GaAs processing methods* (Artech House, Boston, **1990**).

[120] A. Ichimiya and P. I. Cohen, *Reflection High Energy Electron Diffraction*, pp. 4–6 (Cambridge University Press, 1, **2004**).

[121] A. Bernard, E. Delamarche, H. Schmid and B. Michel, "Printing Patterns of Proteins," *Langmuir*, vol. 14, pp. 2225–2229, **1998**.

[122] Y. Xia and G. Whitesides, "Soft Lithography," *Annual Reviews in Materials Science*, vol. 28, pp. 153–184, **1998**.

[123] S. Chou, P. Krauss and P. Renstrom, "Imprint of sub-25 nm vias and trenches in polymers," *Applied Physics Letters*, vol. 67, pp. 3114–3116, **1995**.

[124] G. Jung, E. Johnston-Halperin, W. Wu, Z. Yu and S. Wang, "Circuit fabrication at 17 nm half-pitch by nanoimprint lithography," *Nano Letters*, vol. 6, pp. 351–354, **2006**.

[125] S. Rivzi, *Handbook of Photomask Manufacturing Technology*, pp. 327–335 (CRC Press, London, **2005**).

[126] R. Williams, *Modern GaAs processing methods*, pp. 120–125 (Artech House, Boston, **1990**).

[127] A. Wong, "Microlithography: Trends, Challenges, Solutions, and Their Impact on Design," *IEEE Micro*, pp. 12–21, **2003**.

[128] K. Mistry, C. Allen, C. Auth, B. Beattie and D. Bergstrom, "A 45nm Logic Technology with High-k+ Metal Gate Transistors, Strained Silicon, 9 Cu Interconnect Layers, 193nm Dry Patterning, and 100% Pb-free Packaging," in *IEEE International Electron Devices Meeting* (**2007**).

[129] M. Levenson, N. Viswanathan and R. Simpson, "Improving resolution in photolithography with a phase-shifting mask," *Electron Devices*, vol. 29, pp. 1828–1836, **1982**.

[130] M. Ono, M. Saito, T. Yoshitomi, C. Fiegna, T. Ohguro and H. Iwai, "A 40 nm gate length n-MOSFET," *IEEE Transactions on Electron Devices*, vol. 42, pp. 1822 – 1830, **1995**.

[131] J. Horstmann, U. Hilleringmann and K. Goser, "Matching analysis of deposition defined 50-nm MOSFET's," *IEEE Transactions on Electron Devices*, vol. 45, pp. 299 – 306, **1998**.

[132] K. Kallis, J. Horstmann, A. Wiggershaus and H. Fiedler, "Manufacturing considerations of lithography independent nano-MOS-transisters in the sub-25 nm-region," in *International Conference on Solid-State and Integrated Circuit Technology*, pp. 55 – 57 (**2006**).

[133] J. Bjorkholm, "EUV Lithography—The Successor to Optical Lithography?" *Intel Technology Journal*, pp. 1–8, **1998**.

[134] H. Solak, C. David, J. Gobrecht and V. Golovkina, "Sub-50 nm period patterns with EUV interference lithography," *Microelectronic Engineering*, vol. 67-68, pp. 56–62, **2003**.

[135] P. Silverman, "Extreme ultraviolet lithography: overview and development status," *Journal of Microlithography*, vol. 4, pp. 011 006–1 – 011 006–5, **2005**.

[136] A. Biswas and S. Brueck, "Simulation of the 45-nm half-pitch node with 193-nm immersion lithography—imaging interferometric lithography and dipole illumination," *Journal of Microlithography*, vol. 3, pp. 35–43, **2004**.

[137] J. Mulkens, D. Flagello, B. Streefkerk and P. Graeupner, "Benefits and limitations of immersion lithography," *Journal of Microlithography*, vol. 3, pp. 104–114, **2004**.

[138] S. Natarajan, M. Armstrong, M. Bost and R. Brain, "A 32nm logic technology featuring 2nd-generation high-k+ metal-gate transistors, enhanced channel Enhanced Channel Strain and 0.171μm^2 SRAM Cell Size in a 291Mb Array," in *IEEE International Electron Devices Meeting* (**2008**).

[139] M. Smayling and V. Axelrad, "32nm and below logic patterning using optimized illumination and double patterning," in *Proceedings of SPIE*, vol. 7274 (**2009**).

[140] A. Broers, "Resolution Limits for Electron-Beam Lithography," *IBM Journal of Research and Development*, vol. 32, pp. 502–513, **1988**.

[141] A. Tseng, K. Chen, C. Chen and K. Ma, "Electron beam lithography in nanoscale fabrication: recent development," *Electronics Packaging Manufacturing*, vol. 26, pp. 141–149, **2003**.

[142] Vistec Semiconductor Systems, "Vectorbeam Series Operators Manual," , **2003**.

[143] G. Owen and P. Rissman, "Proximity effect correction for electron beam lithography by equalization of background dose," *Journal of Applied Physics*, vol. 54, pp. 3573–3581, **1983**.

[144] T. Chang, "Proximity effect in electron-beam lithography," *Journal of Vacuum Science and Technology B: Microelectronics and Nanometer Structures*, vol. 12, pp. 1271–1275, **1975**.

[145] H. Eisenmann, T. Waas and H. Hartmann, "PROXECCO—Proximity effect correction by convolution," *Journal of Vacuum Science and Technology B: Microelectronics and Nanometer Structures*, vol. 11, pp. 2741–2745, **1993**.

[146] M. Parikh, "Self-consistent proximity effect correction technique for resist exposure (SPECTRE)," *Journal of Vacuum Science and Technology B: Microelectronics and Nanometer Structures*, vol. 15, pp. 931–933, **1978**.

[147] D. Hasko, S. Yasin and A. Mumtaz, "Influence of developer and development conditions on the behavior of high molecular weight electron beam resists," *Journal of Vacuum Science and Technology B: Microelectronics and Nanometer Structures*, vol. 18, pp. 3441–3444, **2000**.

[148] J. Greeneich, "Developer Characteristics of Poly(Methyl Methacrylate) Electron Resist," *Journal of The Electrochemical Society*, vol. 122, pp. 970–976, **1975**.

[149] M. Khoury and D. Ferry, "Effect of molecular weight on poly (methyl methacrylate) resolution," *Journal of Vacuum Science and Technology B: Microelectronics and Nanometer Structures*, vol. 14, pp. 75–79, **1996**.

[150] S. Thoms, D. Macintyre and M. McCarthy, "Sub - 35 nm metal gratings fabricated using PMMA with high contrast developers," *Microelectronic Engineering*, vol. 41-42, pp. 207–210, **1998**.

[151] A. Endoh, Y. Yamashita, K. Shinohara, M. Higashiwaki, K. Hikosaka, T. Mimura, S. Hiyamizu and T. Matsui, *Fabrication technology and device performance of sub-50-nm-gate InP-based HEMTs*, pp. 448–451 (**2001**).

[152] B. Lim, H. Lee, D. Shin, S. Kim, H. Park and J. Rhee, "Sub-100 nm T-gate fabrication using a positive resist ZEP520/P(MMA-MAA)/PMMA trilayer by double exposure at 50 kV e-beam lithography," *Materials Science in Semiconductor Processing*, vol. 7, pp. 7–11, **2004**.

[153] T. Yamaguchi and H. Namatsu, "Effect of developer molecular size on roughness of dissolution front in electron-beam resist," *Journal of Vacuum Science and Technology B: Microelectronics and Nanometer Structures*, vol. 22, pp. 1037–1043, **2004**.

[154] K. Kurihara, K. Iwadate, H. Namatsu, M. Nagase, H. Takenaka and K. Murase, "An electron beam nanolithography system and its application to

Si nanofabrication," *Japanese Journal of Applied Physics*, vol. 34, pp. 6940–6946, **1995**.

[155] K. Lister, S. Thoms, D. Macintyre, C. Wilkinson and J. Weaver, "Direct imprint of sub-10 nm features into metal using diamond and SiC stamps," *Journal of Vacuum Science and Technology B: Microelectronics and Nanometer Structures*, vol. 22, pp. 3257–3259, **2004**.

[156] K. Lister, B. Casey, P. Dobson, S. Thoms, D. Macintyre, C. Wilkinson and J. Weaver, "Pattern transfer of a 23 nm-period grating and sub-15 nm dots into CVD diamond," *Microelectronic Engineering*, vol. 73-74, pp. 319–322, **2004**.

[157] L. Mollard, G. Cunge, S. Tedesco, B. Dal'zotto and J. Foucher, "HSQ hybrid lithography for 20 nm CMOS devices development," *Microelectronic Engineering*, vol. 61-62, pp. 755–761, **2002**.

[158] O. Nalamasu, M. Cheng, J. Kometani and S. Vaidya, "Development of a chemically amplified positive resist material for single-layer deep-UV lithography," in *Proceedings of SPIE*, vol. 1262, pp. 32–48 (**1990**).

[159] Y. Chen, D. Macintyre and S. Thoms, "Electron beam lithography process for T- and Gamma-shaped gate fabrication using chemically amplified DUV resists and PMMA," *Journal of Vacuum Science and Technology B: Microelectronics and Nanometer Structures*, vol. 17, p. 2507, **1999**.

[160] S. Thoms, S. Beaumont, C. Wilkinson and J. Frost, "Ultrasmall device fabrication using dry etching of GaAs," *Microelectronic Engineering*, pp. 249–256, **1986**.

[161] S. Murad, M. Rahman, N. Johnson, S. Thoms, S. Beaumont and C. Wilkinson, "Dry etching damage in III–V semiconductors," *Journal of Vacuum Science and Technology B: Microelectronics and Nanometer Structures*, vol. 14, pp. 3658–3662, **1996**.

[162] L. Pantisano, A. Paccagnelle and P. Colombo, "Plasma damage impact on nMOS electrical characteristics during a CCS stress," *Plasma Process-Induced Damage*, pp. 73–76, **1999**.

[163] K. L. Seaward, N. J. Moll and W. F. Stickle, "Surface contamination and damage from CF_4 and SF_6 reactive ion etching of silicon oxide on gallium arsenide," *Journal of Electronic Materials*, vol. 19, pp. 385–391, **1990**.

[164] R. Williams, *Modern GaAs Processing Methods*, pp. 95–101 (Artech House, Boston, **1990**).

[165] K. Williams, "Etch Rates for Micromachining Processing," *Journal of Microelectromechanical Systems*, vol. 5, pp. 256–269, **1996**.

[166] R. Williams, *Modern GaAs Processing Methods*, pp. 153–155 (Artech House, Boston, **1990**).

[167] A. Grill, *Cold plasma in Materials Fabrication*, pp. 35–39 (IEEE Press, New York, **1994**).

[168] —, *Cold plasma in Materials Fabrication*, pp. 13–17 (IEEE Press, New York, **1994**).

[169] —, *Cold plasma in Materials Fabrication*, pp. 219–222 (IEEE Press, New York, **1994**).

[170] —, *Cold plasma in Materials Fabrication*, pp. 94–111 (IEEE Press, New York, **1994**).

[171] —, *Cold plasma in Materials Fabrication*, pp. 223–241 (IEEE Press, New York, **1994**).

[172] R. Williams, *Modern GaAs Processing Methods*, pp. 161–168 (Artech House, Boston, **1990**).

[173] A. Grill, *Cold Plasma in Materials Fabrication*, pp. 195–201 (IEEE Press, New York, **1994**).

[174] G. Ponchak, L. Katehi, N. Center and O. Cleveland, "Characteristics of finite ground coplanar waveguide lumped elements," in *IEEE/MTT-S International Microwave Symposium* (**1997**).

[175] D. Moran, E. Boyd, H. McLelland, K. E. Y. Chen, D. Macintyre, S. Thoms, C. Stanley and I. Thayne, "Novel technologies for the realisation of GaAs pHEMTs with 120 nm self-aligned and nanoimprinted T-gates," *Microelectronic Engineering*, vol. 67-68, p. 769, **2003**.

[176] U. Mishra, A. Brown, L. Jelloian, M. Thompson, L. Nguyen and S. Rosenbaum, "Novel High Performance Self-Aligned 0.15 micron Long T-Gate AlInAs-GaInAs HEMTs," in *IEEE International Electron Devices Meeting*, pp. 101–104 (**1989**).

[177] D. Moran, E. Boyd, F. McEwan, H. McLelland, C. Stanley and I. Thayne, "Sub 100nm T-gate uniformity in InP HEMT technology," in *Proceedings of GaAs ManTech* (**2004**).

[178] M. Shur, *Introduction to Electronic Devices*, pp. 172–178 (John Wiley & Sons, New York, **1996**).

[179] L. V. der Pauw, "A Method of Measuring the Resistivity and Hall Coefficient of Lamellae of Arbitrary Shape," *Philips Technical Review*, vol. 20, pp. 220–224, **1959**.

[180] A. Goetzberger, R. Scarlett and W. Shockley, "Research and Investigation of Inverse Epitaxial UHF Power Transistors," *oai.dtic.mil*, **1964**.

[181] M. Lijadi, F. Pardo, N. Bardou and J. Pelouard, "Floating contact transmission line modelling: An improved method for ohmic contact resistance ...," *Solid State Electronics*, vol. 49, pp. 1655–1661, **2005**.

[182] H. Murrmann and D. Widmann, "Current crowding on metal contacts to planar devices," *IEEE Transactions on Electron Devices*, vol. 16, pp. 1022–1024, **1969**.

[183] G. Reeves and H. Harrison, "Obtaining the specific contact resistance from transmission line model measurements," *Electron Device Letters*, vol. 3, p. 111, **1982**.

[184] D. Moran, *Self-aligned Short Gate Length III-V HEMT Technology*, Phd thesis, University of Glasgow, **2003**.

[185] D. Anderson, L. Smith and J. Gruszynski, "S-Parameter Techniques : HP Test and Measurement Application Note 95-1," , **1996**.

[186] M. Berroth and R. Bosch, "Broad-band determination of the FET small-signal equivalent circuit," *IEEE Transactions on Microwave Theory and Techniques*, vol. 38, p. 891, **1990**.

[187] D. Edgar, K. Elgaid, F. Williamson, S. Ferguson, A. Ross, F. Doherty, I. Thayne, M. Taylor and S. Beaumont, "W-Band Performance of Coplanar Waveguide on Thinned Substrates," in *European Microwave Conference*, vol. 3, pp. 363 – 366 (**1999**).

[188] G. Dambrine, A. Cappy, F. Heliodore and E. Playez, "A new method for determining the FET small-signal equivalent circuit," *IEEE Transactions on Microwave Theory and Techniques*, vol. 36, pp. 1151–1159, **1988**.

[189] W. Curtice and R. Camisa, "Self-Consistent GaAs FET Models for Amplifier Design and Device Diagnostics," *IEEE Transactions on Microwave Theory and Techniques*, vol. 32, p. 1573, **1984**.

[190] S. Sze, "Physics of Semiconductor Devices," *John Wiley & Sons*, vol. New York, p. 162, **1981**.

[191] J. Rollett, "Stability and Power-Gain Invariants of Linear Twoports," *IRE Transactions on Circuit Theory*, vol. 9, pp. 29 – 32, **1962**.

[192] H. Takahashi, F. Murai, S. Asai and H. Kodera, "Reproducible submicron gate fabrication of GaAs FET by plasma etching," in *IEEE International Electron Devices Meeting*, vol. 22, pp. 214–217 (**1976**).

[193] S. Takahashi, F. Murai and H. Kodera, "Submicrometer gate fabrication of GaAs MESFET by plasma etching," *IEEE Transactions on Electron Devices*, vol. 25, pp. 1213–1218, **1978**.

[194] H. Morkoc, J. Andrews, R. Sankaran and J. Dully, "Tungsten/gold gate GaAs microwave f.e.t," *Electronics Letters*, vol. 14, pp. 514–515, **1978**.

[195] Y. Todokoro, "Double-layer resist films for submicrometer electron-beam lithography," *IEEE Transactions on Electron Devices*, vol. 27, pp. 1443–1448, **1980**.

[196] M. Matsumura, K. Tsutsui and Y. Naruke, "Submicrometre lift-off line with T-shaped cross-sectional form," *Electronics Letters*, vol. 17, pp. 429–430, **1981**.

[197] P. Chao, P. Smith, S. Wanuga, J. Hwang, W. Perkins, R. Tiberio and E. Wolf, "Electron-beam fabrication of quarter-micron T-shaped-gate FETs using a new tri-layer resist system," in *IEEE International Electron Devices Meeting*, vol. 29, pp. 613–616 (**1983**).

[198] A. Tessmann, "220-GHz metamorphic HEMT amplifier MMICs for high-resolution imaging applications," *IEEE Journal of Solid-State Circuits*, vol. 40, pp. 2070–2076, **2005**.

[199] R. Lai, P. Huang, R. Grundbacher, D. Farkas, A. Cavus, P. Liu, P. Chin, Y. Chou, M. Barsky, R. Tsai, R. Raja and A. Oki, "0.07 μm InP HEMT MMIC Technology for G-band Power Amplifiers," in *International Conference on Indium Phosphide and Related Materials*, pp. 39–41 (**2006**).

[200] D. Dawson, L. Samoska, A. Fung, K. Lee and R. Lai, "Beyond G-Band: A 235 GHz InP MMIC Amplifier," *IEEE Microwave and Wireless Components Letters*, vol. 15, pp. 874–876, **2005**.

[201] L. Ocola, D. Tennant and P. Ye, "Bilayer process for T-gates and Gamma-gates using 100kV e-beam lithography," *Microelectronic Engineering*, vol. 67, pp. 104–108, **2003**.

[202] K. Elgaid, H. McLelland, C. Stariley and I. Thayne, "Low noise W-band MMMIC amplifier using 50nm InP technology for millimeterwave receivers applications," in *International Conference on Indium Phosphide and Related Materials*, p. 523 (**2005**).

[203] Y. Chen, D. Macintyre, X. Cao, E. Boyd, D. Moran, H. McLelland, M. Holland, C. Stanley, I. Thayne and S. Thoms, "Fabrication of ultrashort T gates using a PMMA/LOR/UVIII resist stack," *Journal of Vacuum Science and Technology B: Microelectronics and Nanometer Structures*, vol. 21, pp. 3012–3016, **2003**.

[204] Y. Chen, K. Peng and Z. Cui, "Fabrication of ultra-short T gates by a two-step electron beam lithography process," *Microelectronic Engineering*, vol. 73-74, pp. 662–665, **2004**.

[205] Y. Chen, D. Edgar, X. Li, D. Macintyre and S. Thoms, "Fabrication of 30 nm T gates using SiN as a supporting and definition layer," *Journal of Vacuum Science and Technology B: Microelectronics and Nanometer Structures*, vol. 18, p. 3521, **2000**.

[206] T. Suemitsu, T. Ishii, H. Yokoyama, Y. Umeda, T. Enoki, Y. Ishii and T. Tamamura, "30-nm-gate InAlAs/InGaAs HEMTs lattice-matched to InP substrates," in *IEEE International Electron Devices Meeting*, p. 223 (**1998**).

[207] T. Ishii, H. Nozawa and T. Tamamura, "Nanocomposite resist system," *Applied Physics Letters*, vol. 70, pp. 1110–1112, **1997**.

[208] Y. Yamashita, A. Endoh, K. Shinohara, K. Hikosaka, T. Matsui, S. Hiyamizu and T. Mimura, "Pseudomorphic $In_{0.52}Al_{0.48}As/In_{0.7}Ga_{0.3}As$ HEMTs with an ultrahigh f_t of 562 GHz," *Electron Device Letters*, vol. 23, p. 573, **2002**.

[209] S. Kwang-Seok and K. Dae-Hyun, "Nanometer scale InGaAs HEMT technology for ultra high speed IC," in *International Conference on Indium Phosphide and Related Materials*, pp. 30–35 (**2006**).

[210] S.-J. Yeon, J. Lee and S. Kwang-Seok, "Gate Length Reduction Technology for Pseudomorphic $In_{0.52}Al_{0.48}As/In_{0.7}Ga_{0.3}As$ High Electron Mobility Transistors," *Japanese Journal of Applied Physics*, vol. 46, pp. 2296–2299, **2007**.

[211] P. Smith, S. Liu, M. Kao, P. Ho, S. Wang, K. Duh, S. Fu and P. Chao, "W-band high efficiency InP-based power HEMT with 600 GHz f_{max}," *IEEE Microwave and Guided Wave Letters*, vol. 5, pp. 230–232, **1995**.

[212] R. Lai, X. Mei, W. Deal, W. Yoshida, Y. Kim, P. Liu, J. Lee, J. Uyeda, V. Radisic, M. Lange, T. Gaier, L. Samoska and A. Fung, "Sub 50 nm InP HEMT Device with Fmax Greater than 1 THz Sub 50 nm InP HEMT Device with Fmax Greater than 1 THz," in *IEEE International Electron Devices Meeting*, pp. 609–611 (**2007**).

[213] Y. Chen, G. Wang, W. Schaff, P. Tasker, K. Kavanagh and L. Eastman, "A high performance 0.12 um T-shape gate $Ga_{0.5}In_{0.5}As/Al_{0.5}In_{0.5}As$ MODFET grown by MBE lattice-mismatched on a GaAs substrate," in *IEEE International Electron Devices Meeting*, pp. 431–434 (**1987**).

[214] G. Wang, Y. Chen, W. Schaff and L. Eastman, "A 0.1-μm gate $Al_{0.5}In_{0.5}As/Ga_{0.5}In_{0.5}As$ MODFET fabricated on GaAs substrates," *IEEE Transactions on Electron Devices*, vol. 35, pp. 818–823, **1988**.

[215] G. Ng, K. Radhakrishnan and H. Wang, "Are We There Yet?-A Metamorphic HEMT and HBT Perspective," in *European Gallium Arsenide and Other Semiconductor Symposium*, pp. 13–19 (**2005**).

[216] J. Chang, J. Chen, J. Fernandez and H. Wieder, "Strain relaxation of compositionally graded InGaAs buffer layers for modulation-doped InGaAs/InAlAs heterostructures," *Applied Physics Letters*, vol. 60, pp. 1129–1131, **1992**.

[217] K. Yuan and K. Radhakrishnan, "High breakdown voltage $In_{0.52}Al_{0.48}As/In_{0.53}Ga_{0.47}As$ metamorphic HEMT using $In_xGa_{1-x}P$ Graded Buffer," in *International Conference on Indium Phosphide and Related Materials*, pp. 161–164 (**2002**).

[218] W. Hoke, P. Lemonias, J. Mosca, P. Lyman, A. Torabi, P. Marsh, R. McTaggart, S. Lardizabal and K. Hetzler, "Molecular beam epitaxial growth and device performance of metamorphic high electron mobility transistor structures fabricated on GaAs substrates," in *North American Molecular Beam Epitaxy Conference*, vol. 17, p. 1131 (**1999**).

[219] M. Behet, "Metamorphic InGaAs/InAlAs quantum well structures grown on GaAs substrates for high electron mobility transistor applications," *Applied Physics Letters*, vol. 73, pp. 2760–2762, **1998**.

[220] K. van der Zanden, M. Behet and G. Borghs, "Comparison of Metamorphic InGaAs/InAlAs HEMT's on GaAs with InP based LM HEMT's," in *International Conference on Gallium Arsenide Manufacturing Technology* (**1999**).

[221] M. Haupt, K. Kohler, P. Ganser, S. Emminger, S. Müller and W. Rothemund, "Growth of high quality $Al_{0.48}In_{0.52}As/Ga_{0.47}In_{0.53}As$ heterostructures using strain relaxed $Al_xGa_yIn_{1-x-y}As$ buffer layers on GaAs," *Applied Physics Letters*, vol. 69, pp. 412–414, **1996**.

[222] D. Dumka, H. Tserng, M. Kao and E. B. III, "High-performance double-recessed enhancement-mode metamorphic HEMTs on 4-in GaAs substrates," *Electron Device Letters*, vol. 24, pp. 135–137, **2003**.

[223] M. Kao, E. B. III, T. Yun, C. Campbell and M. Heins, "Metamorphic HEMT technology for millimeter-wave and 40-Gb/s fiber-optics applications," in *International Conference on Indium Phosphide and Related Materials*, pp. 361–364 (**2003**).

[224] K. Elgaid, D. Moran, H. McLelland, M. Holland and I. Thayne, "Low noise high performance 50nm T-gate metamorphic HEMT with cut-off frequency f_t of 440 GHz for millimeterwave imaging receivers applications," in *International Conference on Indium Phosphide and Related Materials*, pp. 141–143 (**2005**).

[225] K. Sung-Won, L. Kang-Min, L. Jae-Hak and S. Kwang-Seok, "High-performance 0.1-μm $In_{0.4}AlAs/In_{0.35}GaAs$ MHEMTs with Ar plasma treatment," *Electron Device Letters*, vol. 26, pp. 787–789, **2005**.

[226] S. Mertens, J. del Alamo and C. MIT, "Electrical degradation of InAlAs/InGaAs metamorphic high-electron mobility transistors," in *IEEE International Electron Devices Meeting*, pp. 193–196 (**2001**).

[227] G. Meneghesso and E. Zanoni, "Failure modes and mechanisms of InP-based and metamorphic high electron mobility transistors," *Microelectronics Reliability*, vol. 42, pp. 685–708, **2002**.

[228] R. Menozzi, "Hot electron effects and degradation of GaAs and InP HEMTs for microwave and millimetre-wave applications," *Semiconductor Science and Technology*, vol. 13, p. 1053, **1998**.

[229] Y. Yamashita, A. Endoh, K. Shinohara, M. Higashiwaki, K. Hikosaka, T. Mimura, S. Hiyamizu and T. Matsui, "Ultra-short 25-nm-gate lattice-matched InAlAs/InGaAs HEMTs within the range of 400 GHz cutoff frequency," *IEEE Electron Device Letters*, vol. 22, pp. 367–369, **2001**.

[230] Y. Kwon, M. Tutt, G. Ng, D. Pavlidis, T. Brock, P. Marsh, J. Oh, J. Castagne and N. Linh, "Gate-recess and device geometry impact on the microwave performance and noise properties of 0.1μm InAlAs/InGaAs HEMT's," in *Proceedings IEEE/Cornell Conference on Advanced Concepts in High Speed Semiconductor Devices and Circuits*, pp. 141–150 (**1991**).

[231] D. Xu, H. Heiss, M. Sexl, S. Kraus, G. Bohm, G. Tränkle, G. Weimann and G. Abstreiter, "2 S/mm Transconductance InAs-Inserted-Channel Modulation Doped Field Effect Transistors with a Very Close Gate-to-Channel Separation of 14.5 nm," *Japanese Journal of Applied Physics*, vol. 36, pp. 470–472, **1997**.

[232] T. Suemitsu, T. Enoki, M. Tomizawa, N. Shigekawa and Y. Ishii, "Mechanism and structural dependence of kink phenomena in InAlAs/InGaAs HEMTs," in *International Conference on Indium Phosphide and Related Materials*, pp. 365–368 (**1997**).

[233] K. Shinohara, Y. Yamashita, A. Endoh, I. Watanabe, K. Hikosaka, T. Mimura, S. Hiyamizu and T. Matsui, "Nanogate InP-HEMT technology for ultrahigh-speed performance," in *International Conference on Indium Phosphide and Related Materials*, p. 721 (**2004**).

[234] H. Fourre, F. Diette and A. Cappy, "Selective wet etching of lattice-matched InGaAs/InAlAs on InP and metamorphic InGaAs/InAlAs on GaAs using succinic acid/hydrogen peroxide solution," *Journal of Vacuum Science and Technology B: Microelectronics and Nanometer Structures*, vol. 14, p. 3400, **1996**.

[235] S. Bahl and I. del Alamo, "Elimination of mesa-sidewall gate leakage in InAlAs/InGaAs heterostructures by selective sidewall recessing Elimination of mesa-sidewall gate leakage in InAlAs/InGaAs heterostructures by selective sidewall recessing," *Electron Device Letters*, vol. 13, pp. 195–197, **1992**.

[236] E. Moon, J.-L. Lee and H. Yoo, "Selective wet etching of GaAs on $Al_xGa_{1-x}As$ for AlGaAs/InGaAs/AlGaAs pseudomorphic high electron mobility transistor," *Journal of Applied Physics*, vol. 84, p. 3933, **1998**.

[237] M. Tong, K. Nummila, A. Ketterson, I. Adesida, C. Caneau and R. Bhat, "InAlAs/InGaAs/InP MODFET's with uniform threshold voltage obtained by selective wet gate recess," *Electron Device Letters*, vol. 13, pp. 525–527, **1992**.

[238] G. DeSalvo, C. Bozada, J. Ebel, D. Look, J. Barrette, C. Cerny, R. Dettmer, J. Gillespie, C. Havasy, T. Jenkins, K. Nakano, C. Pettiford, T. Quach, J. Sewell and G. Via, "Wet Chemical Digital Etching of GaAs at Room Temperature," *Journal of The Electrochemical Society*, vol. 143, p. 3652, **1996**.

[239] X. Cao and I. Thayne, "Novel high uniformity highly reproducible non-selective wet digital gate recess etch process for InP HEMTs," *Microelectronic Engineering*, vol. 67-68, p. 333, **2003**.

[240] M. Boudrissa, E. Delos, X. Wallaert, D. Theron and J. D. Jaeger, "A 0.15-μm 60-GHz high-power composite channel GaInAs/InP HEMTwith low gate current," *Electron Device Letters*, vol. 22, pp. 257–259, **2001**.

[241] Y. Liu and H. Wang, "Influence of silicon nitride passivation on transport properties in InAlAs/InGaAs/InP composite channel high electron mobility transistor structures," *Journal of Vacuum Science and Technology B: Microelectronics and Nanometer Structures*, vol. 24, pp. 1711–1715, **2006**.

[242] Y. Chen, P. Chin, D. Ingram, R. Lai and R. Grundbacher, "Composite-channel InP HEMT for W-band power amplifiers," in *International Conference on Indium Phosphide and Related Materials*, pp. 305–306 (**1999**).

[243] K. Yu, H. Chuang, K. Lin, S. Cheng and C. Cheng, "Improved temperature-dependent performances of a novel InGaP-InGaAs-GaAs double channel pseudomorphic high electron transistor (DC-PHEMT)," *Electron Devices*, vol. 49, pp. 1687–1693, **2002**.

[244] H. Maher, J. Decobert, A. Falcou, M. L. Pallec, G. Post, Y. Nissim and A. Scavennec, "A triple channel HEMT on InP (Camel HEMT) for large-signal high-speed applications," *IEEE Transactions on Electron Devices*, vol. 46, pp. 32–37, **1999**.

[245] A. Leuther, R. Weber, M. Dammann and M. Schlechtweg, "Metamorphic 50 nm InAs-channel HEMT," in *International Conference on Indium Phosphide and Related Materials*, pp. 129–132 (**2005**).

[246] D. Kim and J. del Alamo, "Logic Performance of 40 nm InAs HEMTs," in *IEEE International Electron Devices Meeting*, pp. 629–632 (**2007**).

[247] M. Borg, E. Lefebvre, M. Malmkvist, L. Desplanque, L. Desplanque, X. Wallart, Y. Roelens, G. Dambrine, A. Cappy, S. Bollaert and J. Grahn, "Effect of gate length in InAs/AlSb HEMTs biased for low power or high gain," *Solid State Electronics*, **2008**.

[248] W. Deal, R. Tsai, M. Lange, J. Boos and B. Bennett, "A W-band InAs/AlSb low-noise/low-power amplifier," *Microwave and Wireless Components Letters*, vol. 15, pp. 208–210, **2005**.

[249] Y. Chou, M. Lange, B. Bennett, J. Boos and J. Yang, "0.1 μm $In_{0.2}Al_{0.8}Sb - InAs$ HEMT Low-Noise Amplifiers for Ultralow-Power Applications," in *IEEE International Electron Devices Meeting*, pp. 617–620 (**2007**).

[250] J. del Alamo and D.-H. Kim, "Beyond CMOS: Logic Suitability of InGaAs HEMTs," in *International Conference on Indium Phosphide and Related Materials*, pp. 51–54 (**2007**).

[251] T. Ashley, L. Buckle, M. Emeny, M. Fearn, D. Hayes, K. Hilton, R. Jefferies, T. Martin, T. Phillips, J. Powell, A. Tang, D. Wallis and P. Wilding, "Indium Antimonide based Quantum Well FETs for Ultra-high Frequency, Low Power Dissipation Circuits," in *European Microwave Integrated Circuits Conference, 2006*, pp. 29 – 30 (**2006**).

[252] T. Ashley, L. Buckle, S. Datta, M. Emeny, D. Hayes, K. Hilton, R. Jefferies, T. Martin, T. Phillips, D. Wallis, P. Wilding and R. Chau, "Heterogeneous InSb quantum well transistors on silicon for ultra-high speed, low power logic applications," *Electronics Letters*, vol. 43, **2007**.

[253] T. Ashley, A. Barnes, L. Buckle, S. Datta, A. Dean, M. Emery, M. Fearn, D. Hayes, K. Hilton, R. Jefferies, T. Martin, K. Nash, T. Phillips, W. Tang, P. Wilding and R. Chau, "Novel InSb-based quantum well transistors for ultra-high speed, low power logic applications," in *International Conference on Solid-State and Integrated Circuit Technology*, vol. 3, pp. 2253 – 2256 vol.3 (**2004**).

[254] M. Radosavljevic, T. Ashley, A. Andreev, S. Coomber, G. Dewey, M. Emeny, M. Fearn, D. Hayes, K. Hilton, M. Hudait, R. Jefferies, T. Martin, R. Pillarisetty, W. Rachmady, T. Rakshit, S. Smith, M. Uren, D. Wallis, P. Wilding and R. Chau, "High-performance 40nm gate length InSb p-channel compressively strained quantum well field effect transistors for low-power (VCC=0.5V) logic applications," in *IEEE International Electron Devices Meeting*, pp. 1–4 (**2008**).

[255] T. Horiuchi, T. Homma, Y. Murao and K. Okumura, "An asymmetric sidewall process for high performance LDD MOSFET's," *IEEE Transactions on Electron Devices*, vol. 41, pp. 186–190, **1994**.

[256] X. Li, R. Hill, H. Zhou, C. Wilkinson and I. Thayne, "A low damage Si_3N_4 sidewall spacer process for self-aligned sub-100nm III–V MOSFETs," *Microelectronic Engineering*, vol. 85, pp. 996–999, **2008**.

[257] L. Nguyen, A. Brown, M. Thompson and L. Jelloian, "50-nm self-aligned-gate pseudomorphic AlInAs/GaInAs high electron mobility transistors," *IEEE Transactions on Electron Devices*, vol. 39, p. 2007, **1992**.

[258] D. Moran, E. Boyd, K. Elgaid, F. McEwan, H. McLelland, C. Stanley and I. Thayne, "Self-aligned T-gate InP HEMT realisation through double delta doping and a non-annealed ohmic process," *Microelectronic Engineering*, vol. 73-74, p. 814, **2004**.

[259] Y. Nakasha, Y. Kawano, M. Sato, T. Takahashi and K. Hamaguchi, "Ultra high-speed and ultra low-noise InP HEMTs," *Fujitsu Scientific and Technical Journal*, vol. 43, pp. 486–494, **2007**.

[260] J. Mateos, T. González, D. Pardo, V. Hoel and A. Cappy, "Effect of the T-gate on the performance of recessed HEMTs. A Monte Carlo analysis," *Semiconductor Science and Technology*, vol. 14, pp. 864–870, **1999**.

[261] J. Lopez, T. Gonzalez, D. Pardo, S. Bollaert, T. Parenty and A. Cappy, "Design Optimization of AlInAs–GaInAs HEMTs for High-Frequency Applications," *IEEE Transactions on Electron Devices*, vol. 51, pp. 521–528, **2004**.

[262] K. Kalna, A. Asenov, K. Elgaid and I. Thayne, "Performance of aggressively scaled pseudomorphic HEMTs: a Monte Carlo simulation study," in *Third International Euroconference on Advanced Semiconductor Devices*, pp. 55–58 (**2000**).

[263] —, "Scaling of pHEMTs to decanano dimensions," in *VLSI design*, vol. 13, p. 155 (**2000**).

[264] S. Babiker, A. Asenov, N. Cameron, S. Beaumont and J. Barker, "Complete Monte Carlo RF analysis of "real" short-channel compound FET's," *IEEE Transactions on Electron Devices*, vol. 45, p. 1644, **1998**.

[265] K. Hur and R. Compton, "Airbridged-gate MESFETs fabricated by isotropic reactive ion etching," *IEEE Transactions on Electron Devices*, vol. 40, p. 1736, **1993**.

[266] P. Chao, M. Shur, R. Tiberio, K. Duh, P. Smith, J. Ballingall, P. Ho and A. Jabra, "DC and microwave characteristics of sub-0.1 um gate-length planar-doped pseudomorphic HEMTs," *IEEE Transactions on Electron Devices*, vol. 36, p. 461, **1989**.

[267] Y. Chan, D. Pavlidis and G. Ng, "The influence of gate-feeder/mesa-edge contacting on sidegating effects in $In_{0.52}Al_{0.48}As/In_{0.53}Ga_{0.47}As$ heterostructure FET's," *Electron Device Letters*, vol. 12, p. 360, **1991**.

[268] G. Truitt, D. Heston and J. Klein, "A New Low-Noise FET Structure," *IEEE Transactions on Microwave Theory and Techniques*, vol. 38, pp. 1944–1948, **1990**.

[269] R. Lai, M. Barsky, T. Huang, M. Sholley and H. Wang, "An InP HEMT MMIC LNA with 7.2-dB gain at 190 GHz," *Microwave and Guided Wave Letters*, vol. 8, pp. 393–395, **1998**.

[270] C. Liechti, "Microwave Field-Effect Transistors-1976," *Microwave Theory and Techniques*, vol. 24, pp. 279–300, **1976**.

[271] M. Passlack, M. Hong, E. Schubert and J. Kwo, "*In situ* fabricated GaO–GaAs structures with low interface recombination velocity," *Applied Physics Letters*, vol. 66, pp. 625–627, **1995**.

[272] M. Passlack, J. Abrokwah, R. Droopad and Z. Yu, "Self-aligned GaAs p-channel enhancement mode MOS heterostructurefield-effect transistor," *IEEE Electron Device Letters*, vol. 23, pp. 508–510, **2002**.

[273] D. Kim and J. D. Alamo, "Lateral and Vertical Scaling of $In_{0.7}Ga_{0.3}As$ HEMTs for Post-Si-CMOS Logic Applications," *Electron Devices*, vol. 55, pp. 2546–2553, **2008**.

[274] D.-H. Kim and J. del Alamo, "Impact of lateral engineering on the logic performance of sub-50 nm InGaAs HEMTs," in *International Semiconductor Device Research Symposium*, pp. 1 – 2 (**2007**).

[275] T. Suemitsu and M. Tokumitsu, "InP HEMT Technology for High-Speed Logic and Communications," *IEICE Transactions on Electronics*, vol. E90-C, pp. 917–922, **2007**.

[276] K. Shinohara, W. Ha, M. Rodwell and B. Brar, "Extremely High $g_m>$ 2.2 S/mm and $f_t>$ 550 GHz in 30-nm Enhancement-Mode InP-HEMTs with Pt/Mo/Ti/Pt/Au Buried Gate," in *International Conference on Indium Phosphide and Related Materials*, pp. 18–21 (**2007**).

[277] K. Chen, T. Enoki, K. Maezawa, K. Arai and M. Yamamoto, "High-performance InP-based enhancement-mode HEMTs using non-alloyed ohmic contacts and Pt-based buried-gate technologies," *Electron Devices*, vol. 43, pp. 252–257, **1996**.

[278] M. Fischetti and S. Laux, "Monte Carlo simulation of transport in technologically significantsemiconductors of the diamond and zinc-blende structures," *IEEE Transactions on Electron Devices*, vol. 38, pp. 650–660, **1991**.

[279] D. Park and K. Brennan, "Theoretical analysis of an $Al_{0.15}Ga_{0.85}As/In_{0.15}Ga_{0.85}As$ pseudomorphic HEMT using an ensemble Monte Carlo simulation," *IEEE Transactions on Electron Devices*, vol. 36, pp. 1254–1263, **1989**.

[280] N. Pilgrim, W. Batty and R. Kelsall, "Electrothermal Monte Carlo Simulations of InGaAs/AlGaAs HEMTs," *Journal of Computational Electronics*, vol. 2, pp. 207–211, **2003**.

[281] N. Seoane, A. Garcia-Loureiro, K. Kalna and A. Asenov, "Impact of intrinsic parameter fluctuations on the performance of HEMTs studied with a 3D parallel drift-diffusion simulator," *Solid State Electronics*, vol. 51, pp. 481–488, **2007**.

[282] R. Quay, K. Hess, R. Reuter, M. Schlechtweg and T. Grave, "Nonlinear electronic transport and device performance of HEMTs," *Electron Devices*, vol. 48, pp. 210–217, **2001**.

[283] I. Kizilyalli, M. Artaki, N. Shah and A. Chandra, "Scaling properties and short-channel effects in submicrometer AlGaAs/GaAs MODFET's: A Monte Carlo study," *IEEE Transactions on Electron Devices*, vol. 40, pp. 234 – 249, **1993**.

[284] S. Babiker, A. Asenov, N. Cameron and S. Beaumont, "Simple approach to include external resistances in the Monte Carlo simulation of MESFETs and HEMTs," *IEEE Transactions on Electron Devices*, vol. 43, p. 2032, **1996**.

[285] F. Medjdoub, M. Zaknoune, X. Wallart, C. Gaquiere, F. Dessenne, J. Thobel and D. Theron, "InP HEMT downscaling for power applications at W band," *IEEE Transactions on Electron Devices*, vol. 52, pp. 2136–2143, **2005**.

[286] A. Endoh, Y. Yamashita, K. Shinohara, K. Hikosaka, T. Matsui, S. Hiyamizu and T. Mimura, "InP-Based High Electron Mobility Transistors with a Very Short Gate-Channel Distance," *Japanese Journal of Applied Physics*, vol. 42, pp. 2214–2218, **2003**.

[287] R. Singh and C. Snowden, "Small-signal characterization of microwave and millimeter-waveHEMT's based on a physical model," *IEEE Transactions on Microwave Theory and Techniques*, vol. 44, pp. 114–121, **1996**.

[288] C. Morton, J. Atherton, C. Snowden and R. Pollard, "A large-signal physical HEMT model," in *IEEE MTT-S International Microwave Symposium Digest*, pp. 1759–1762 (**1996**).

[289] R. Drury and C. Snowden, "A quasi-two-dimensional HEMT model for microwave CAD applications," *IEEE Transactions on Electron Devices*, vol. 42, pp. 1026–1032, **1995**.

[290] P. Fay, M. Arafa, W. Wohlmuth and C. Caneau, "Design , fabrication, and performance of high-speed monolithically integrated InAlAs/InGaAs/InP MSM/HEMT photoreceivers," *Journal of Lightwave Technology*, vol. 15, pp. 1871–1879, **1997**.

[291] K. Kalna and A. Asenov, "Role of multiple delta doping in PHEMTs scaled to sub-100 nm dimensions," *Solid-State Electronics*, vol. 48, p. 1223, **2004**.

[292] —, "Nonequilibrium transport in scaled high electron mobility transistors," *Semiconductor Science and Technology*, vol. 17, pp. 579–584, **2002**.

[293] J. Adams, I. Thayne, M. Taylor, C. Wilkinson, S. Beaumont, N. Johnson, A. Kean and C. Stanley, "Very high frequency performance of nanometre scale GaAs MESFETs," *IEEE Transactions on Electron Devices*, vol. 36, p. 2612, **1989**.

[294] I. Thayne, K. Elgaid, M. Taylor, M. Holland, S. Fairbairn, N. Cameron, S. Beaumont and G. Belle, "Low-frequency noise of selectively dry-etch gate-recessed GaAs MESFETs," *Electronics Letters*, vol. 31, pp. 324 – 326, **1995**.

[295] P. Chao, P. Smith, S. Palmateer and J. Hwang, "Electron-beam fabrication of GaAs low-noise MESFET's using a new trilayer resist technique," *IEEE Transactions on Electron Devices*, vol. 32, pp. 1042 – 1046, **1985**.

[296] I. Thayne, M. Holland, Y. Chen, W. Li, A. Paulsen, S. Beaumont and P. Bhattacharya, "Comparison of 80-200 nm gate length $Al_{0.25}GaAs/GaAs/(GaAs:AlAs)$, $Al_{0.3}GaAs/In_{0.15}GaAs/GaAs$, and $In_{0.52}AlAs/In_{0.65}GaAs/InP$ HEMTs," in *IEEE International Electron Devices Meeting*, pp. 225 – 228 (**1993**).

[297] Y. Chen, D. Macintyre and S. Thoms, "Fabrication of T-shaped gates using UVIII chemically amplified DUV resist and PMMA," *Electronics Letters*, vol. 35, pp. 338–339, **1999**.

[298] Y. Chen, T. Lodhi, H. McLelland, D. Edgar, D. Macintyre, S. Thoms, C. Stanley and I. Thayne, "First demonstration of InAlAs/InGaAs HEMTs using T-gates fabricated by a bilayer of UVIII and PMMA resists," in *IEEE International Symposium on High Performance Electron Devices for Microwave and Optoelectronic Applications*, pp. 202–205 (**2000**).

[299] Y. Chen, D. Macintyre and S. Thoms, "Effect of resist sensitivity ratio on T-gate fabrication," in *International conference on electron, ion, and photon beam technology and nanofabrication*, vol. 19, p. 2494 (**2001**).

[300] X. Cao, E. Boyd, H. McLelland, S. Thoms, C. Stanley and I. Thayne, "mm-wave Performance of 50nm T-Gate AlGaAs/InGaAs pseudomorphic High Electron Mobility Transistors with f_t of 200 GHz," in *European Microwave Week - GAAS Symposium* (**2003**).

[301] X. Cao, I. Thayne, S. Thoms, M. Holland and C. Stanley, "High Performance 50nm T-Gate $In_{0.52}Al_{0.48}As/In_{0.53}Ga_{0.47}As$ Metamorphic High Electron Mobility Transistors," in *GaAs Symposium*, pp. 197–199 (**2003**).

[302] X. Cao, S. Thoms, D. Macintyre, H. McLelland, E. Boyd, K. Elgaid, R. Hill, C. Stanley and I. Thayne, "Fabrication and performance of 50 nm T-gates for InP high electron mobility transistors," *Microelectronic Engineering*, vol. 73-74, p. 818, **2004**.

[303] E. Boyd, H. Zhou, H. McLelland, D. Moran and S. Thoms, "Fabrication of 30nm T-Gate High Electron Mobility Transistors Using a Bi-Layer of PMMA and UVIII," *Optoelectronic and Microelectronic Materials and Devices*, pp. 25–28, **2004**.

[304] E. Boyd, *Development of Advanced Technologies for the Fabrication of III-V High Electron Mobility Transistors.*, Phd thesis, University of Glasgow, **2004**.

[305] D. Drouin, A. Couture, D. Joly, X. Tastet, V. Aimez and R. Gauvin, "CASINO V2. 42-a fast and easy-to-use modeling tool for scanning electron microscopy and microanalysis users," *Scanning*, vol. 29, pp. 92–101, **2007**.

[306] J. Lee, J. Shim, H. Yoon and H. Kim, "Fabrication Technology and Device Performance of SiN Assisted 0.15 μm Gate $In_{0.52}Al_{0.48}As/In_{0.53}Ga_{0.47}As$ Metamorphic HEMT on GaAs Substrate," *Journal of the Korean Physical Society*, vol. 42, pp. S662–665, **2003**.

[307] H. Wenchuang, G. Bernstein, K. Sarveswaran and M. Lieberman, "Low temperature development of PMMA for sub-10-nm electron beam lithography," in *IEEE Conference on Nanotechnology*, vol. 2, pp. 602–605 vol. 2 (**2003**).

[308] W. Hu, K. Sarveswaran, M. Lieberman and G. Bernstein, "Sub-10 nm electron beam lithography using cold development of poly(methylmethacrylate)," *Journal of Vacuum Science and Technology B: Microelectronics and Nanometer Structures*, vol. 22, pp. 1711–1716, **2004**.

[309] Z. Lu and A. Cartwright, "Reduction of metal linewidths through a combination of low-temperature and ultrasonic development of poly(methylmethacrylate) using electron-beam lithography," in *Nanoengineering: Fabrication, Properties, Optics, and Devices III*, vol. 6327, pp. 63 270O–11 (**2006**).

[310] W. Chen and H. Ahmed, "Fabrication of sub-10 nm structures by lift-off and by etching after electron-beam exposure of poly(methylmethacrylate) resist on solid substrates," in *International conference on electron, ion, and photon beam technology and nanofabrication*, vol. 11, pp. 2519–2523 (**1993**).

[311] B. Maile, W. Henschel, H. Kurz, B. Rienks and R. Polman, "Sub-10 nm Linewidth and Overlay Performance Achieved with a Fine-Tuned EBPG-5000 TFE Electron Beam Lithography System," *Japanese Journal of Applied Physics*, vol. 39, pp. 6836–6842, **2000**.

[312] N. Jin, S. Choi, L. Wang, G. Chen, D. Kim, V. Kumar and I. Adesida, "Nanometer-scale gaps in hydrogen silsesquioxane resist for T-gate fabrication," *Journal of Vacuum Science and Technology B: Microelectronics and Nanometer Structures*, vol. 25, p. 2081, **2007**.

[313] K. Elgaid, H. McLelland, S. Ferguson, X. Cao and E. Boyd, "An array-based design methodology for the realisation of 94GHz MMMIC amplifiers," in *European Microwave Conference*, pp. 13–15 (**2004**).

[314] H. Zhou, K. Elgaid, C. Wilkinson and I. Thayne, "Low-hydrogen-content silicon nitride deposited at room temperature by inductively coupled plasma deposition," *Japanese Journal of Applied Physics*, vol. 45, pp. 8388–8392, **2006**.

[315] C. Canalias, V. Pasiskevicius, R. Clemens and F. Laurell, "Submicron periodically poled flux-grown KTiOPO," *Applied Physics Letters*, **2003**.

[316] M. Urech, *Spin-dependant transport in lateral nano-devices based on magnetic tunnel junctions*, Phd thesis, Kungliga Tekniska Högskolan, **2006**.

[317] Shipley, "Microposit Remover 1165 Datasheet," p. 3, **2008**.

[318] D. Moran, K. Kalna, E. Boyd, F. McEwan, H. McLelland, L. Zhuang, C. Stanley, A. Asenov and I. Thayne, "Self-aligned 0.12 μm T-gate $In_{.53}Ga_{.47}As/In_{.52}Al_{.48}As$ HEMT technology utilising a non-annealed ohmic contact strategy," in *ESSDERC Proceedings*, pp. 315–318 (**2003**).

[319] D. Moran, E. Boyd, K. Elgaid, H. McLelland, C. Stanley and I. Thayne, "50nm T-gate lattice-matched InP HEMTs with f_t of 430GHz using a non-annealed ohmic contact process," in *GaAs Symposium*, pp. 311–314 (**2004**).

[320] K. Elgaid, X. Li, F. Williamson, H. McLelland, S. Ferguson, M. Holland, S. Beaumont and I. Thayne, "Optimisation of DC and RF performance of GaAs HEMT-based Schottky diodes," *Electronics Letters*, vol. 35, p. 1678, **1999**.

[321] K. J. Choi, J.-L. Lee and H. M. Yoo, "Effects of deep levels on transconductance dispersion in AlGaAs/InGaAs pseudomorphic high electron mobility transistor," *Applied Physics Letters*, vol. 75, p. 1580, **1999**.

[322] V. R. Balakrishnan, V. Kumar and S. Ghosh, "The origin of low-frequency negative transconductance dispersion in a pseudomorphic HEMT," *Semiconductor Science and Technology*, vol. 20, pp. 783–787, **2005**.

[323] K. J. Choi and J.-L. Lee, "Interpretation of transconductance dispersion in GaAs MESFET using deep level transient spectroscopy," *IEEE Transactions on Electron Devices*, vol. 48, pp. 190–195, **2001**.

[324] T. Enoki, H. Ito and Y. Ishii, "Reliability study on InAlAs/InGaAs HEMTs with an InP recess-etch stopper and refractory gate metal," *Solid-State Electronics*, vol. 41, pp. 1651–1656, **1997**.

[325] R. Iyer and D. Lile, "Role of polysulfides in the passivation of the InP surface," *Applied Physics Letters*, vol. 59, pp. 437–439, **1991**.

[326] R. Iyer, R. Chang, A. Dubey and D. Lile, "The effect of phosphorous and sulfur treatment on the surface properties of InP," *Journal of Vacuum Science and Technology B: Microelectronics and Nanometer Structures*, vol. 6, pp. 1174–1179, **1988**.

[327] H. Hasegawa, "Controlled formation of high Schottky barriers on InP and related materials," in *International Conference on Indium Phosphide and Related Materials*, pp. 451–454 (**1998**).

[328] J. Leclercq, E. Bergignat and G. Hollinger, "Surface chemistry of InAlAs after $(NH_4)_2S_x$ sulphidation," *Semiconductor Science and Technology*, vol. 10, pp. 95–100, **1995**.

[329] J.-L. Lee, D. Kim, S. Maeng, H. Park, J. Kang and Y. Lee, "Improvement of breakdown characteristics of a GaAs power field-effect transistor using $(NH_4)_2S_x$ treatment," *Journal of Applied Physics*, vol. 73, pp. 3539–3542, **1993**.

[330] S. Shikata, H. Okada and H. Hayashi, "Suppression of the emitter size effect on the current gain of AlGaAs/GaAs heterojunction bipolar transistor by utilizing $(NH_4)_2S_x$ treatment," *Journal of Applied Physics*, vol. 69, pp. 2717–2718, **1991**.

[331] C. Sandroff, R. Nottenburg, J. Bischoff and R. Bhat, "Dramatic enhancement in the gain of a GaAs/AlGaAs heterostructure bipolar transistor by surface chemical passivation," *Applied Physics Letters*, vol. 51, pp. 33–35, **1987**.

[332] H. Uchiyama, T. Taniguchi and M. Kudo, "Control of plasma induced fluorine damage in P-HEMT using InSb barrier layer," in *International Conference on Indium Phosphide and Related Materials*, pp. 727–730 (**2004**).

[333] A. Taguchi, T. Ohno and T. Sasaki, "Fluorine atoms in AlAs, GaAs, and InAs: Stable state, diffusion, and carrier passivation," *Physical Review B*, vol. 62, p. 1821, **2000**.

[334] M. Dammann, M. Chertouk, W. Jantz, K. Kohler, K. Schmidt and G. Weimann, "Reliability of passivated 0.15 μm InAlAs-InGaAs HEMTs with pseudomorphic channel," in *IEEE International Reliability Physics Symposium*, pp. 99–102 (**1999**).

[335] Y. Yamamoto, N. Hayafuji, N. Fujii, K. Kadoiwa, N. Yoshida, T. Sonoda, S. Takamiya and S. Mitsui, "Donor passivation in n-AlInAs layers by fluorine," in *International Conference on Indium Phosphide and Related Materials*, pp. 265–268 (**1995**).

[336] K. Docherty, S. Thoms, P. Dobson and J. Weaver, "Improvements to the alignment process in a commercial vector scan electron beam lithography tool," *Microelectronic Engineering*, vol. 85, pp. 761–763, **2008**.

[337] S. Bentley, X. Li, D. Moran and I. Thayne, "Fabrication of 22nm T-gates for HEMT applications," *Microelectronic Engineering*, vol. 85, pp. 1375–1378, **2008**.

[338] H. Namatsu, M. Nagase, K. Kurihara, K. Iwadate, T. Furuta and K. Murase, "Fabrication of sub-10-nm silicon lines with minimum fluctuation," *Journal of Vacuum Science and Technology B: Microelectronics and Nanometer Structures*, vol. 13, pp. 1473–1476, **1995**.

[339] Y. Choi, N. Lindert, P. Xuan, S. Tang and D. Ha, "Sub-20nm CMOS FinFET Technologies," in *IEEE International Electron Devices Meeting*, pp. 421–424 (**2001**).

[340] Y. Choi, T. King and C. Hu, "A spacer patterning technology for nanoscale CMOS," *IEEE Transactions on Electron Devices*, vol. 49, pp. 436–441, **2002**.

[341] J. Chung, M. Jeng, J. Moon, A. Wu, T. Chan and P. Ko, "Deep-submicrometer MOS device fabrication using a photoresist-ashing technique," *IEEE Electron Device Letters*, vol. 9, pp. 186–188, **1988**.

[342] X. Li, H. Zhou, J. Abrokwah, P. Zurcher, K. Rajagopalan, W. Liu, R. Gregory, M. Passlack and I. Thayne, "Low damage ashing and etching processes for ion implanted resist and Si_3N_4 removal by ICP and RIE methods," *Microelectronic Engineering*, vol. 85, pp. 966–968, **2008**.

[343] "Front End Processes," in *International Technology Roadmap for Semiconductors, 2007 Edition*, pp. 1–65 (**2007**).

[344] S. Rishton and D. Kern, "Point exposure distribution measurements for proximity correction in electron beam lithography on a sub-100 nm scale," *Journal of Vacuum Science and Technology B: Microelectronics and Nanometer Structures*, vol. 5, pp. 135–141, **1987**.

[345] E. Anderson, D. Olynick, W. Chao, B. Harteneck and E. Veklerov, "Influence of sub-100 nm scattering on high-energy electron beam lithography," *Journal of Vacuum Science*, vol. 19, pp. 2504–2507.

[346] T. Broekaert and C. Fonstad, "Novel, Organic Acid-Based Etchants for InGaAlAs/InP Heterostructure Devices witjh AlAs Etch-Stop Layers," *Journal of The Electrochemical Society*, vol. 139, pp. 2306–2309, **1992**.

[347] M. Malmkvist, S. Wang and J. Grahn, "Epitaxial Optimization of 130-nm Gate-Length InGaAs/InAlAs/InP HEMTs for High-Frequency Applications," *IEEE Transactions on Electron Devices*, vol. 55, pp. 268–275, **2008**.

[348] A. Endoh, Y. Yamashita, K. Shinohara, M. Higashiwaki, K. Hikosaka, T. Matsui, S. Hiyamizu and T. Mimura, "InP HEMTs: physics, applications, and future," in *International Conference on Indium Phosphide and Related Materials*, p. 5 (**2003**).

[349] Y. Awano, M. Kosugi, S. Kuroda, T. Mimura and M. Abe, "Electron dynamics and device physics of short-channel HEMTs: transverse-domain formation, velocity overshoot, and short-channel effects," in *IEEE/Cornell Conference on Advanced Concepts in High Speed Semiconductor Devices and Circuits*, p. 46 (**1989**).

[350] S. Pearton, "Ion implantation for isolation of III-V semiconductors." *Materials Science Reports*, vol. 4, p. 367, **1990**.

[351] S. Ahmed, R. Gwilliam and B. Sealy, "Proton implantation for isolation of n-type GaAs layers at different substrate temperatures," *Semiconductor Science and Technology*, vol. 16, pp. L28–L31, **2001**.

[352] W. Duncan and S. Matteson, "Compensation in n-type GaAs resulting from nitrogen ion implantation," *Journal of Applied Physics*, vol. 56, pp. 1059–1062, **1984**.

[353] S. Pearton, C. Abernathy, M. Panish, R. Hamm and L. Lunardi, "Implant-induced high-resistivity regions in InP and InGaAs," *Journal of Applied Physics*, vol. 66, p. 656, **1989**.

[354] S. Pantelides, *Deep Centers in Semiconductors : A State-of-the-Art Approach* (**1992**).

[355] P. Too, S. Ahmed, R. Gwilliam and B. Sealy, "Electrical isolation of InP and InGaAs using iron and krypton," *Electronics Letters*, vol. 40, pp. 1302–1304, **2004**.

[356] J. Ziegler, "The Stopping and Range of Ions in Matter," *www.srim.org*, **2005**.

[357] K. Elgaid, H. McLelland, M. Holland, D. Moran, C. Stanley and I. Thayne, "50-nm T-Gate Metamorphic GaAs HEMTs With f_t of 440 GHz and Noise Figure of 0.7 dB at 26 GHz," *Electron Device Letters*, vol. 26, p. 784, **2005**.