The Development Of Planar High-K/III-V P-Channel MOSFETs For Post-Silicon CMOS

A Thesis Submitted To
School Of Engineering
Electronics and Nanoscale Engineering Research Division
University Of Glasgow
In Fulfilment Of The Requirements
For The Degree Of
Doctor Of Philosophy

By
Uthayasankaran Peralagu, *M.Eng. (Hons)*
June 2016

Abstract

Conventional Si complementary-metal-oxide-semiconductor (CMOS) scaling is fast approaching its limits. The extension of the logic device roadmap for future enhancements in transistor performance requires non-Si materials and new device architectures. III-V materials, due to their superior electron transport properties, are well poised to replace Si as the channel material beyond the 10 nm technology node to mitigate the performance loss of Si transistors from further reductions in supply voltage to minimise power dissipation in logic circuits. However several key challenges, including a high quality dielectric/III-V gate stack, a low-resistance source/drain (S/D) technology, heterointegration onto a Si platform and a viable III-V p-metal-oxide-semiconductor field-effect-transistor (MOSFET), need to be addressed before III-Vs can be employed in CMOS.

This Thesis specifically addressed the development and demonstration of planar III-V p-MOSFETs, to complement the n-MOSFET, thereby enabling an all III-V CMOS technology to be realised. This work explored the application of InGaAs and InGaSb material systems as the channel, in conjunction with Al_2O_3/metal gate stacks, for p-MOSFET development based on the buried-channel flatband device architecture. The body of work undertaken comprised material development, process module development and integration into a robust fabrication flow for the demonstration of p-channel devices.

The parameter space in the design of the device layer structure, based around the III-V channel/barrier material options of $In_{x\geq 0.53}Ga_{1-x}As/In_{0.52}Al_{0.48}As$ and $In_{x\geq 0.1}Ga_{1-x}Sb/AlSb$, was systematically examined to improve hole channel transport. A mobility of $433\,cm^2/Vs$, the highest room temperature hole mobility of any InGaAs quantum-well channel reported to date, was obtained for the $In_{0.85}Ga_{0.15}As$ (2.1% strain) structure.

S/D ohmic contacts were developed based on thermally annealed Au/Zn/Au metallisation and validated using transmission line model test structures. The effects of metallisation

thickness, diffusion barriers and de-oxidation conditions were examined. Contacts to InGaSb-channel structures were found to be sensitive to de-oxidation conditions.

A fabrication process, based on a lithographically-aligned double ohmic patterning approach, was realised for deep submicron gate-to-source/drain gap (L_{side}) scaling to minimise the access resistance, thereby mitigating the effects of parasitic S/D series resistance on transistor performance. The developed process yielded gaps as small as 20 nm.

For high-k integration on GaSb, ex-situ ammonium sulphide ($(NH_4)_2S$) treatments, in the range 1%-22%, for 10 min at 295 K were systematically explored for improving the electrical properties of the Al_2O_3/GaSb interface. Electrical and physical characterisation indicated the 1% treatment to be most effective with interface trap densities in the range of $4-10\times10^{12}\,\mathrm{cm^{-2}eV^{-1}}$ in the lower half of the bandgap. An extended study, comprising additional immersion times at each sulphide concentration, was further undertaken to determine the surface roughness and the etching nature of the treatments on GaSb.

A number of p-MOSFETs based on III-V-channels with the most promising hole transport and integration of the developed process modules were successfully demonstrated in this work. Although the non-inverted InGaAs-channel devices showed good current modulation and switch-off characteristics, several aspects of performance were non-ideal; depletion-mode operation, modest drive current ($I_{\mathrm{d,sat}}$=1.14 mA/mm), double peaked transconductance (g_{m}=1.06 mS/mm), high subthreshold swing (SS=301 mV/dec) and high on-resistance (R_{on}=845 k$\Omega.\mu$m). Despite demonstrating substantial improvement in the on-state metrics of $I_{\mathrm{d,sat}}$ (11×), g_{m} (5.5×) and R_{on} (5.6×), inverted devices did not switch-off. Scaling gate-to-source/drain gap (L_{side}) from 1 μm down to 70 nm improved $I_{\mathrm{d,sat}}$ (72.4 mA/mm) by a factor of 3.6 and g_{m} (25.8 mS/mm) by a factor of 4.1 in inverted InGaAs-channel devices. Well-controlled current modulation and good saturation behaviour was observed for InGaSb-channel devices. In the on-state $In_{0.3}Ga_{0.7}Sb$-channel ($I_{\mathrm{d,sat}}$=49.4 mA/mm, g_{m}=12.3 mS/mm, R_{on}=31.7 k$\Omega.\mu$m) and $In_{0.4}Ga_{0.6}Sb$-channel ($I_{\mathrm{d,sat}}$=38 mA/mm, g_{m}=11.9 mS/mm, R_{on}=73.5 k$\Omega.\mu$m) devices outperformed the InGaAs-channel devices. However the devices could not be switched off. These findings indicate that III-V p-MOSFETs based on InGaSb as opposed to InGaAs channels are more suited as the p-channel option for post-Si CMOS.

Acknowledgements

This Thesis has been a long time coming and its completion marks a major milestone in my life. I am indebted to several people for their input to this work and for helping me in one way or another to make my PhD journey a meaningful one.

First and foremost, I would like to express my heartfelt gratitude to my supervisor Professor Iain Thayne for presenting me with the opportunity to undertake this project and for being an excellent mentor. Thank you Iain for your guidance and support when needed, and also for allowing me the freedom to pursue my own research direction and to learn from my mistakes. I would also like to thank Iain for his compassion and understanding during a trying personal experience. My second supervisor Professor Douglas Paul also deserves thanks for his knowledge and help with simulations.

I would like to thank Dr. Kristel Fobelets, of Imperial College London, and Dr. Khaled Elgaid, of University of Glasgow, for serving as my PhD examiners. Many thanks for your constructive feedback during the viva. Dr. Phil Dobson, of University of Glasgow, also deserves thanks for acting as the convener during the viva proceedings.

All the present and former members of the Ultrafast Systems Group and the III-V MOSFET project, at the University of Glasgow, deserve thanks for sharing their knowledge and allowing me the opportunity to utilise their expertise. The following people deserve special mention. I would like to thank Dr. Martin Holland for growing high-quality InGaAs p-type epi-layers and for useful discussions on how to improve the channel transport. Dr. Paolo Longo deserves thanks for performing TEM analysis on some of my samples early on in the project. Dr. Gary Paterson deserves credit for helping me with temperature-dependent magnetotransport measurements, introducing me to LabVIEW and for his staggering knowledge on semiconductor physics. I would also like to thank Dr. Karol Kalna, Dr. Jason Ayubi-Moak and Dr. Brahim Benbakhti for performing some simulations when required. Dr. Steven Bentley is also due thanks for making me

understand and appreciate device fabrication a lot more, teaching me how to make DC measurements and for the very many discussions in the penultimate year of my PhD. Dr. Xu Li also deserves thanks for all things processing related.

I would also like to thank a number of our external collaborators without whom this work would not have come to fruition. Many thanks to Dr. Paul Hurley, Dr. Ian Povey, Dr. Patrick Carolan and Jun Lin, of Tyndall National Institute, for providing me with much needed ALD capability, temperature dependent measurements and TEM analysis on my antimonide devices. I would also like to thank Professor Ravi Droopad and Dr. Rocios Contreras-Guerrero, of Texas State University, for growing high quality InGaSb p-type epi-layers. Thank you Dr. Richard Hill and Dr. Joe Bennett, of SUNY Poly SEMATECH, for the SIMS offer and analysis. I would like to thank Professor Paul McIntyre and Dr. Jaesoo Ahn, of Stanford University, for ALD dielectrics on my InGaAs devices.

The technical staff of the James Watt Nanofabrication Centre (JWNC) also deserve thanks for running the facility in a smooth and efficient manner. Special mention goes to Helen McLelland, Susan Ferguson and Thomas Reilly for their help with fabrication and for their amazing banter.

Many colleagues and friends, too many to name, also made the PhD experience a worthwhile one. Two in particular are deserving of special mention. Thank you Dr. Kirsty Walls for pushing me to get my Thesis written-up and for proofreading it. Dr. Salah Sharabi deserves thanks for the many coffee breaks, many hours of troubleshooting equipment problems, measurement knowledge and for the discussions at length on all things in life. Thanks Salah for also being a great flatmate.

I would like to thank my girlfriend, Irene Suarez Marina, for her constant words of encouragement and support when I was feeling like this was a never ending journey. Thank you for being there.

Finally, heartfelt thanks to the ones who brought me into this world, my parents. Thank you for your unconditional love, for being supportive of whatever I wanted to pursue in my life and for making me the person I am today. A big thank you to my sister and my niece for all the love and care you show me in the world. All that I have achieved is only worth it because of my family.

Contents

List of Figures

List of Tables

List of Acronyms

Acronym	Description
AC	alternating current
AFM	atomic force microscopy
ALD	atomic layer deposition
ART	aspect ratio trapping
BEF	beam error feedback
BELLE	beam-writer exposure layout for lithographic engineers
BEOL	backend-of-the-line
BG	bandgap
BSE	back scatter electron
BTBT	band-to-band tunnelling
CAD	computer aided design
CB	conduction band
CBO	conduction band offset
CC	constant-current
CCD	charge coupled device
CET	capacitive equivalent thickness
CLT	critical layer thickness
CME	capacitance modulation efficiency
CMOS	complementary-metal-oxide-semiconductor
CVD	chemical vapour deposition
C–V	capacitance-voltage
DAC	digital-to-analog converter
DC	direct current
DIBL	drain-induced barrier lowering
DIW	de-ionised water

DOS	density of states
DUV	deep ultraviolet
DWB	direct wafer bonding
D-mode	depletion-mode
e-beam	electron beam
EBE	electron beam evaporation
EBL	electron beam lithography
EDP	energy-delay product
EOC	electron-optical column
EOT	equivalent oxide thickness
EOT_{corr}	equivalent oxide thickness correction
EUV	extreme ultraviolet
EWF	effective work function
E-mode	enhancement-mode
FE	field emission
FEOL	frontend-of-the-line
FET	field-effect-transistor
FG	forming gas
FGA	forming gas anneal
FLP	Fermi level pinning
GF	gate-first
GL	gate-last
GPIB	general purpose interface bus
GSWU	guard switch unit
HEMT	high electron mobility transistor
HF	high-frequency
HFET	heterostructure field-effect-transistor
HH	heavy-hole
HHMT	high hole mobility transistor
HKMG	high-k/metal gate
HSQ	hydrogen silsesquioxane
IC	integrated circuit
ICP	inductively-coupled plasma
ICP-CVD	inductively-coupled plasma chemical vapour deposition
IL	interfacial layer
IPA	isopropyl alcohol

IPL	interface passivation layer
ITC	Intelligent Temperature Controller
ITRS	International Technology Roadmap for Semiconductors
I–V	current-voltage
JWNC	James Watt Nanofabrication Centre
J–V	current density-voltage
LA	lithographically-aligned
LE	linear extrapolation
LF	low-frequency
LH	light-hole
LPCVD	low-pressure chemical vapour deposition
MBE	molecular beam epitaxy
MFCMU	multi-frequency capacitance measurement unit
MFP	mean free path
MG	midgap
MIBK	methyl isobutyl ketone
ML	monolayer
MLD	monolayer doping
MOCVD	metal-organic chemical vapour deposition
MOS	metal-oxide-semiconductor
MOSFET	metal-oxide-semiconductor field-effect-transistor
M/S	metal-semiconductor
NA	numerical aperture
NCRC	Nonclassical CMOS Research Center
NID	non-intentionally doped
PDA	post deposition anneal
PEALD	plasma-enhanced atomic layer deposition
PECVD	plasma-enhanced chemical vapour deposition
PMA	post metallisation anneal
PMGI	polydimethylglutarimide
PMMA	poly-methyl methacrylate
PSF	point spread function
PVD	physical vapour deposition
QW	quantum-well
RCS	remote Coulomb scattering
RF	radio frequency

RHEED	reflection high energy electron diffraction
RIE	reactive ion etching
RMS	root-mean-square
RT	room temperature
RTA	rapid thermal annealing
SA	self-aligned
SBH	Schottky barrier height
SCE	short-channel effects
SCUU	SMU CMU unify unit
SEG	selective epitaxial regrowth
SEM	scanning electron microscopy
SI	semi-insulating
SIMS	secondary ion mass spectroscopy
SMU	source/monitor unit
SRC	Semiconductor Research Corporation
S/D	source/drain
TE	thermionic emission
TEM	transmission electron microscopy
TFE	thermionic-field emission
TLM	transmission line model
TMA	trimethyl-aluminum
UHV	ultra-high vacuum
UV	ultraviolet
VB	valence band
VBO	valence band offset
VRU	variable resolution unit
WFLP	weak Fermi level pinning
XPS	x-ray photoemission spectroscopy
1D-PS	one-dimensional self-consistent Poisson-Schrödinger
2DHG	two-dimensional hole gas

List of Symbols

Symbol	Description
B_{z}	transverse magnetic field
C_{ca}	chromatic aberration coefficient
C_{cen}	centroid capacitance
C_{d}	depletion layer capacitance
C_{dep}	depletion capacitance
$C_{\mathrm{d,max}}$	maximum depletion layer capacitance
C_{dos}	density of states (or quantum) capacitance
C_{fb}	flatband capacitance
C_{gc}	gate-to-channel capacitance
C_{il}	interfacial layer capacitance
C_{inv}	inversion capacitance
C_{it}	interface trap capacitance
C_{max}	maximum capacitance
C_{min}	minimum capacitance
C_{ox}	gate dielectric (or oxide) capacitance
C_{s}	semiconductor capacitance
C_{sa}	spherical aberration coefficient
C_{sfb}	semiconductor capacitance at flatband
C_{HF}	high-frequency capacitance
C_{LF}	low-frequency capacitance
d_{c}	chromatic aberration
d_{d}	diffraction limit
d_{g}	beam spot without aberration
d_{min}	final beam spot size
d_{s}	spherical aberration

d_{v}	virtual source size
D	beam exposure dose
D_{it}	interface trap density
e	resist thickness
E_{c}	minimum conduction band energy
E_{g}	energy bandgap of the semiconductor
E_{i}	intrinsic level
E_{maj}	majority carrier band edge energy
E_{ox}	electric field across the gate dielectric
E_{tr}	interface trap energy position
E_{v}	maximum valence band energy
E_{x}	longitudinal electric field
E_{y}	transverse electric field; Hall field
E_{B}	energy barrier height
E_{F}	Fermi level
E_{00}	hole tunnelling probability
F_{L}	Lorentz force
g_{d}	output conductance
g_{m}	(extrinsic) transconductance
g_{mi}	intrinsic transconductance
h	Planck's constant (6.63×10^{-34} J.s)
I_{b}	beam current
I_{d}	drain current
$I_{\mathrm{diffusion}}$	diffusion curent
I_{drift}	drift curent
$I_{\mathrm{d,sat}}$	saturation drain curent
I_{g}	gate current
I_{off}	off-current
I_{on}	on-current
I_{s}	source current
J_{g}	gate leakage current density
J_{FE}	field emission current density
J_{TE}	thermionic emission current density
J_{TFE}	thermionic-field emission current density
k	Boltzmann constant (1.38×10^{-23} J/K)
l_{min}	resolution limit

L_{g}	gate length
L_{side}	gate-to-source/drain separation
L_{T}	transfer length
m_{h}	hole effective mass
m^*	effective carrier mass
M	demagnification
n	channel carrier density
n_{bulk}	bulk hole concentration
n_{i}	intrinsic carrier concentration
n_{s}	surface electron concentration
N_{c}	effective density of states in the conduction band
N_{maj}	density of states in the majority carrier band
N_{v}	effective density of states in the valence band
N_{A}	acceptor doping concentration
p_{chan}	hole density in the channel
p_{s}	surface hole concentration; sheet carrier density
P_{d}	dynamic power
P_{s}	static power
q	elementary charge (1.60×10^{-19} C)
Q_{bulk}	bulk oxide charge
Q_{ch}	channel hole density
Q_{fixed}	fixed oxide charge
Q_{int}	interfacial oxide charge
Q_{inv}	inversion charge density
Q_{it}	interface trap charge
$Q_{\mathrm{m/d}}$	metal/dielectric interface charge
Q_{s}	semiconductor charge
Q_{tot}	total charge
R_{c}	contact resistance
R_{ch}	channel resistance
R_{d}	drain resistance
R_{on}	on-resistance
R_{s}	source resistance
R_{sd}	source/drain series resistance
R_{sh}	semiconductor sheet resistance
R_{side}	gate-to-source/drain (access) resistance

R_{H}	Hall coefficient
s	mask-to-substrate separation
S_{b}	beam step size
SS	subthreshold swing
t_{b}	dwell time
t_{bar}	barrier layer thickness
t_{chan}	channel layer thickness
t_{cen}	inversion layer thickness
t_{ox}	gate dielectric thickness
t_{spa}	spacer layer thickness
T	temperature
V	beam acceleration voltage
V_{bi}	built-in potential
V_{d}	drain voltage
V_{f}	forward bias
V_{fb}	flatband voltage
V_{g}	gate voltage
V_{ox}	potential across the gate dielectric
V_{r}	reverse bias
V_{s}	source voltage
V_{th}	threshold voltage
V_{D}	intrinsic drain voltage
V_{G}	intrinsic gate voltage
V_{H}	Hall voltage
w	barrier width
W_{g}	gate width
x_{d}	depletion layer width
$x_{\mathrm{d,max}}$	maximum depletion width
α	beam convergence angle
γ	resist contrast
ϵ	permittivity of free space (8.85×10^{-14} F/cm)
ϵ_{ox}	relative permittivity of the oxide
ϵ_{s}	relative permittivity of the semiconductor
$\varepsilon_{\shortparallel}$	in-plane biaxial strain
$\varepsilon_{\perp}$	out-of-plane uniaxial strain
κ	scaling factor

λ	electron wavelength
λ_{d}	Debye length
λ_{m}	mean free path
μ	carrier mobility
μ_{eff}	effective mobility
μ_{h}	hole mobility
$\mu_{\mathrm{h,FE}}$	peak field-effect hole mobility
μ_{H}	Hall mobility
ν	Poisson's ratio
ρ	electrical resistivity
ρ_{c}	specific contact resistivity
σ	capture cross section of the trapping state
τ	gate delay
τ_{it}	characteristic trapping lifetime
τ_{L}	minority carrier lifetime
τ_{Ln}	bulk electron lifetime
τ_{Lp}	bulk hole lifetime
τ_{m}	mean free time
τ_{maj}	majority carrier response time
τ_{min}	minority carrier response time
υ_{d}	carrier drift velocity
υ_{inj}	carrier injection velocity
υ_{sat}	saturation velocity
υ_{t}	thermal velocity of carriers
ϕ_{Bp}	Schottky barrier height
ϕ_{m}	metal work function
ϕ_{ms}	metal-semiconductor work function difference
ϕ_{s}	semiconductor work function
χ_{ox}	electron affinity of the oxide
χ_{s}	electron affinity of the semiconductor
ψ_{B}	potential difference between the intrinsic level and the Fermi level
ψ_{s}	surface potential
$\psi_{\mathrm{s,sinv}}$	surface potential at the onset of strong inversion
ψ_{T}	potential difference between the bulk trap level and the intrinsic level
ω	angular frequency
ΔV	beam energy spread

Associated Publications

U. Peralagu, I. M. Povey, P. Carolan, J. Lin, R. Contreras-Guerrero, R. Droopad, P. K. Hurley, and I. G. Thayne, "Electrical and physical characterization of the Al_2O_3/*p*-GaSb interface for 1%, 5%, 10% and 22% $(NH_4)_2S$ surface treatments", *Applied Physics Letters*, vol. 105, no. 16, p. 162907, 2014.

S. W. Chang, X. Li, R. Oxland, S. W. Wang, T. Vasen, C. H. Wang, R. Contreras-Guerrero, K. K. Bhuwalka, G. Doornbos, M. C. Holland, G. Vellianitis, M. J. H. van Dal, B. Duriez, M. Edirisooriya, J. S. Rojas-Ramirez, P. Ramvall, S. Thoms, **U. Peralagu**, C. H. Hsieh, Y. S. Chang, K. M. Yin, L.-E Wernersson, R. Droopad, I. Thayne, M. Passlack, and C. H. Diaz, "InAs N-MOSFETs with record performance of Ion = 600 μA/μm at Ioff = 100 nA/μm (Vd = 0.5 V)", in *IEEE International Electron Devices Meeting (IEDM) Technical Digest*, 2013, pp. 417-420.

U. Peralagu, I. Povey, P. K. Hurley, R. Droopad, and I. G. Thayne, "An investigation of $(NH_4)_2S$ passivation on the electrical, and interfacial properties of the Al_2O_3/GaSb system for p-type and n-type GaSb layers", presented at E-MRS Spring Meeting, Strasbourg, France, May 27-31, 2013.

U. Peralagu, M. C. Holland, G. W. Paterson, and I. G. Thayne, "Strain additivity and its impact on the hole mobility of $In_xGa_{1-x}As$ channels for III-V pMOSFETs", presented at HETECH, Crete, Greece, Oct 20-22, 2010.

U. Peralagu, M. C. Holland, G. W. Paterson, and I. G. Thayne, "The impact of strain engineering on hole mobility of In(x)Ga(1-x)As channels for III-V pMOSFETs", presented as a paper and poster at TECHCON, Austin, Texas, USA, Sep 13-14, 2010.

1

Introduction

1.1 Background and Motivation

The Si metal-oxide-semiconductor field-effect-transistor (MOSFET), invented in 1959 by Kahng and Atalla [1], is the key building block of integrated circuits (ICs) that are found in nearly all modern electronic devices such as mobile phones and computers. In digital applications, which is the mainstream technology of modern microelectronics, MOSFETs serve as an electronic switch. Logic circuits are based on the complementary-metal-oxide-semiconductor (CMOS) technology, which comprises a pair of n- and p-type Si MOSFETs.

"Smaller is better". This motto has been the driving force behind the microelectronics revolution. For over forty years Si CMOS has been advancing along an exponential path of shrinking device dimensions and increasing device density. This trend was first recognised by Gordon Moore in 1965, who predicted that the number of transistors in an IC would double approximately every 2 years [2]. Remarkably, the semiconductor industry has kept pace with Moore's prediction by continuously pushing the limits of technology.

The so-called Moore's law is simply a law of economics. Decreasing the transistor footprint, allows for increased functionality in the same chip area or a similar functionality in a reduced chip area. Either way, the reduction in semiconductor real-estate from downsizing transistors results in a reduced cost per function. But there are other benefits to transistor downsizing. In 1974, Dennard and co-workers [3] showed that by scaling the transistor based on a set of rules, improvements in device performance can be obtained. In this scaling methodology, shown in Table 1.1, the physical dimensions and voltages are reduced by a scaling factor, κ, and the doping is increased by κ, so that the two-dimensional electric field pattern of the transistor is maintained constant. Therefore

Table 1.1: Ideal transistor scaling scheme based on Dennard's scaling laws for a chip of a given size, where κ is the scaling parameter [3].

Device and chip parameters	Scaling factor
Device dimensions: gate length (L_{g}), gate width (W_{g}), oxide thickness (t_{ox}), junction depth (x_{j})	$1/\kappa$
Voltages: supply voltage (V_{d}), threshold voltage (V_{th})	$1/\kappa$
Substrate doping concentration: N_{dop}	κ
Electric field in the device: E_{eff}	1
Gate oxide capacitance: C_{ox}	$1/\kappa$
Drive current in saturation: $I_{\mathrm{d,sat}}$	$1/\kappa$
$I_{\mathrm{d,sat}}$ per unit W_{g}: $I_{\mathrm{d,sat}}/\mu m$	1
Gate delay (switching speed): τ	$1/\kappa$
Power dissipation per device: P	$1/\kappa^2$
Energy-delay product: EDP	$1/\kappa^3$
Number of transistors: N_{T}	κ^2
Clock frequency: f	κ
Power density in chip: P_{chip}	1

as the transistor is made smaller the switching speed is increased by κ and the power dissipation is reduced by κ^2, while the power density of the chip remains unchanged.

Despite Dennard's scaling laws serving as an essential blueprint for transistor scaling, the semiconductor industry did not strictly follow this. Rather, transistor scaling has taken a more aggressive route since the first microprocessor in 1971 [4] as illustrated in Figure 1.1. Over a 30 year period, while the physical dimensions of the transistor decreased by a factor of 100 based on $\kappa{=}\sqrt{2}$ in accordance with Moore's law, the supply voltage only decreased by a factor of 10. As result, the switching speed increased by $1000\times$ as opposed to $100\times$ in the ideal scaling scenario. This additional improvement in device performance came at the cost of chip power consumption, which in ideal scaling is unchanged, increasing by a factor of 10^5. Therefore, power consumption presents the most significant scaling limit.

Power dissipation in CMOS circuits can be categorised into two types; dynamic and static. Dynamic power, P_{d}, is dissipated when switching between logic states while static power, P_{s}, is dissipated when a logic state is maintained between switching events. For over thirty years, the main source of chip power consumption has been P_{d}, given by the product of the on-current and supply voltage ($P_{\mathrm{d}} = I_{\mathrm{d,sat}}V_{\mathrm{d}}$). In 2001, the power dissipated in ICs reached about $100\,\mathrm{W/cm^2}$ [5, 6]. This is the limit that can be tolerated in most practical applications without incurring substantial cooling and packaging costs.

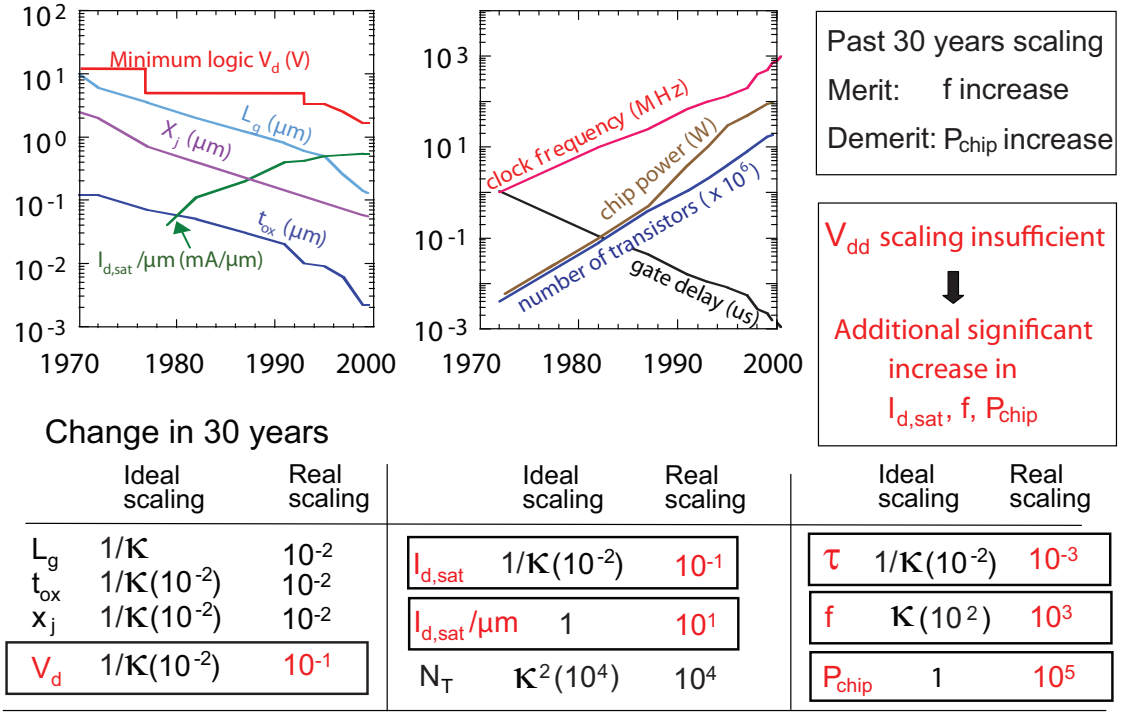


Figure 1.1: Actual industry trend of transistor scaling between the period of 1970 and 2000 [4].

As transistor scaling progresses into the sub-100 nm regime, the once negligible static power appears to be catching up with the dynamic power as observed in Figure 1.2(a). Two effects contribute to the exponential increase in P_s, the first is the fundamental limitation associated with lowering the threshold voltage (V_{th}). Due to the non-scalable characteristic of the thermal voltage, the subthreshold swing (SS), a measure of the steepness of transition between on- and off-state, cannot be scaled beyond the limit of 60 mV/dec at room temperature [5, 6]. As shown in Figure 1.2(b), reducing V_{th} thus results in the off-state leakage current (I_{off}) increasing exponentially. A second effect arises as the transistor gate length is reduced below 100 nm. The reduction in electrostatic gate control over the channel from the stronger influence of the drain manifests as V_{th} roll-off, drain-induced barrier lowering (DIBL) and SS degradation [7]. These effects, collectively known as short-channel effects (SCE), increase the off-state leakage leading to higher static power dissipation. A number of methods including high channel doping, ultra-thin gate oxides and shallow source/drain (S/D) junctions can be used to suppress SCE [7]. These approaches however are accompanied by loss of transistor performance from channel mobility degradation and increased S/D series resistance, and an increase in static power dissipation, from the emergence of gate leakage and band-to-band tunnelling

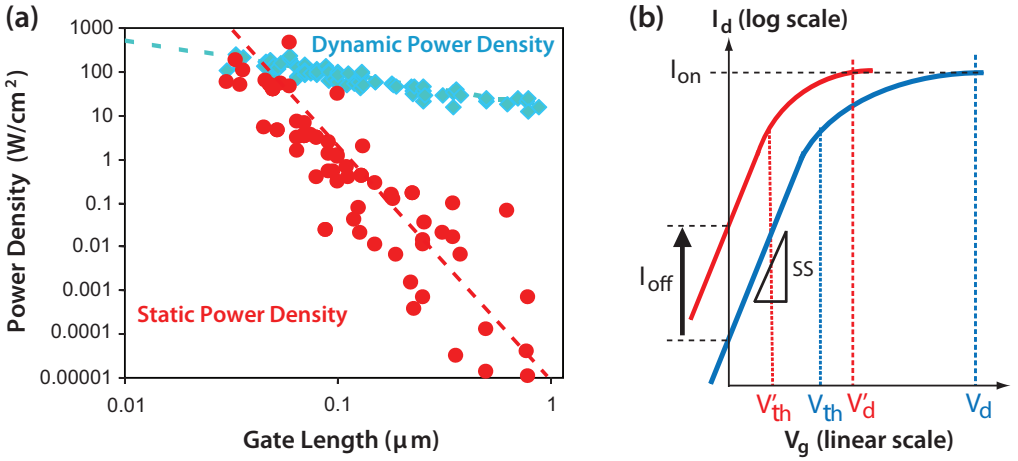


Figure 1.2: (a) Dynamic and static power density trends with gate length scaling [5] and (b) effect of threshold voltage on the off-state leakage current of a MOSFET.

(BTBT) [7].

The mandate for increasing switching speed, at a given overdrive, whilst reducing power dissipation for future scaling is met with conflicting needs between dynamic and static power. Dynamic power pushes for a lower V_{d}, and thus a lower V_{th}, whereas static power demands a higher V_{th}, and hence a higher V_{d}. As a consequence, CMOS has entered into an era of power-constrained scaling where voltage scaling has almost come to a standstill since the 130 nm technology node [8]. To extend Moore's law, new and innovative "technology boosters" have been introduced into conventional Si CMOS. This started from the 90 nm node (2003), where uniaxial compressive and tensile strain was used to enhance the hole and electron mobility in the Si channel. In n-MOSFETs tensile strain was induced by capping the gate stack with a Si_3N_4 layer, while in p-MOSFETs epitaxially regrown SiGe S/D was used to induce compressive strain [9]. By 2007 gate leakage from quantum mechanical tunnelling of carriers through the SiO_2 dielectric layer had increased to intolerable levels as oxide thickness scaled below 2 nm [8]. Further, drive current was degraded from poly-depletion effects associated with the doped polysilicon gate [8]. To mitigate these issues, SiO_2 was replaced with a higher-permittivity (high-k) gate dielectric, HfO_2, of increased physical thickness and CMOS reverted back to using metal gates at the 45 nm node [10]. The introduction of a non-planar 3-dimensional (3D) device architecture at the 22 nm node (2011) signalled the biggest change in CMOS history since high-k/metal gate (HKMG). The improved electrostatic gate control from a Si FinFET, where the gate covers the top and sides of the fin, minimises SCE [11]. Therefore, the FinFET enabled improvements in switching performance and power consumption over the planar equivalent at lower supply voltages [11]. In 2014, the 14 nm node, a second generation Si FinFET,

was released [12]. Presently the industry is priming for the 10 nm node, a third generation Si FinFET due to be released in early 2017.

Beyond the 10 nm node, Si is not expected to feature in CMOS as the channel material due to the unacceptable performance loss with further reductions in supply voltage [13]. In its place, alternate channel materials with higher carrier mobility and velocity will be required [13, 14]. In this regard, III-V compound semiconductors are well placed to supersede Si as the channel material due to their superior electron transport properties. As shown in Figure 1.3, III-Vs demonstrate significantly higher electron mobilities and injection velocities compared to Si.

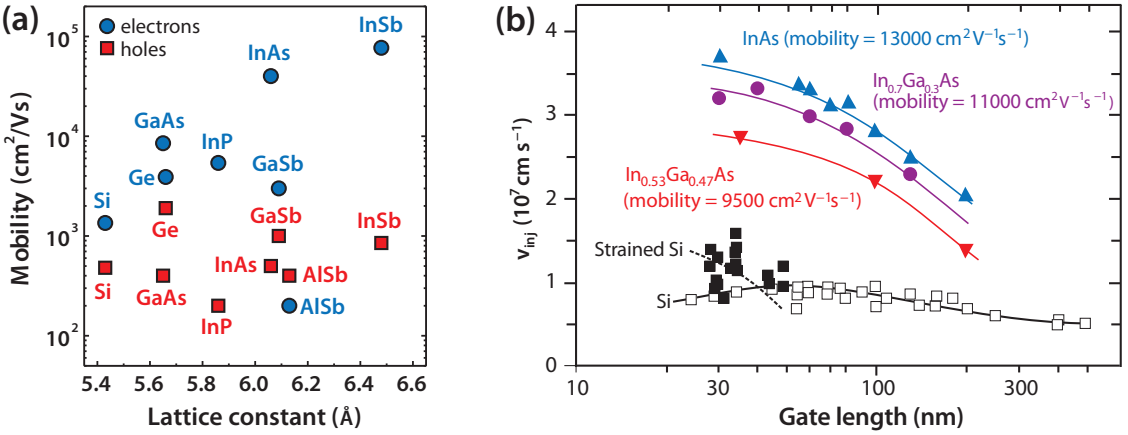


Figure 1.3: (a) Electron and hole mobilities of various semiconductors and (b) electron injection velocity, ν_{inj}, versus gate length for Si n-MOSFETs (at V_{d}=1.1–1.3 V) and III-V HEMTs (at V_{d}=0.5 V) [14].

The benefit of III-V in CMOS can be appreciated by considering its impact on the performance metrics of gate delay and energy-delay product (EDP). The gate delay has an inverse dependence on the average velocity of the carriers, which in the case of short channel devices is limited by the carrier injection velocity (υ_{inj}), as given in Equation (1.1).

$$\tau \propto \frac{V_{\mathrm{d}}}{nq\upsilon_{\mathrm{inj}}}, \tag{1.1}$$

where the product of the injection velocity, channel carrier density, n, and elementary charge, q, determines the drive current of the transistor [15]. The EDP defines the tradeoff between power dissipation and performance and is given as [15]

$$EDP \propto \frac{V_{\mathrm{d}}^3}{nq\upsilon_{\mathrm{inj}}}. \tag{1.2}$$

Figure 1.4 compares the gate delay and EDP, as a function of gate length, between III-V high electron mobility transistors (HEMTs) and Si n-MOSFETs. It is evident that III-Vs demonstrate substantially lower gate delay and EDP as compared to Si at a given gate length. The gain in gate delay for III-V channels saturates with reducing gate length, owing to the increased device parasitics from the non self-aligned (SA) architecture of the HEMT [15]. By adopting a SA technology, it is expected that the performance gain would continue with gate length scaling as indicated by the dashed line in Figure 1.4(a). Based on these merits, there is a strong motivation for III-Vs as a post-Si CMOS option.

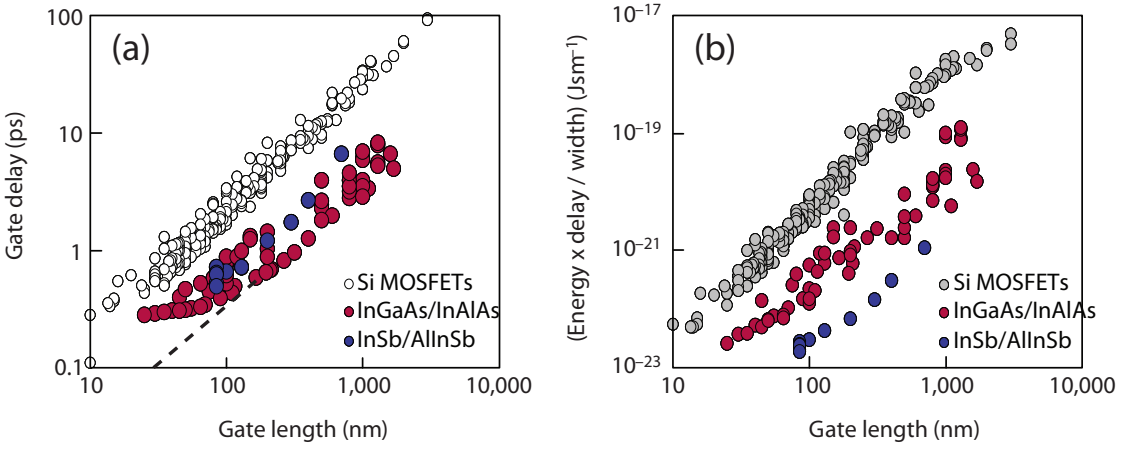


Figure 1.4: Comparison of (a) gate delay and (b) energy-delay product, as a function of gate length, in Si n-MOSFETs and III-V HEMTs [15].

1.2 Key Challenges

The introduction of a new channel material into mainstream CMOS is faced with several obstacles. Before III-Vs can be deployed in future technology nodes, they will have to outperform the Si alternative by a margin of 30-50% [14]. Further, III-V MOSFETs must be manufactured in a cost-effective manner and be scalable over a few technology generations. In order to make this happen several key challenges, as highlighted in Figure 1.5, have to be addressed and are described below.

High Quality Gate stack

Central to the MOSFET is the gate stack. It comprises a metal gate and a dielectric atop the semiconductor channel. A high quality gate stack is crucial for effective gate control of the channel. This requires a suitably scaled dielectric that is free of charge defects, a smooth dielectric/semiconductor interface with minimal interfacial imperfections and high

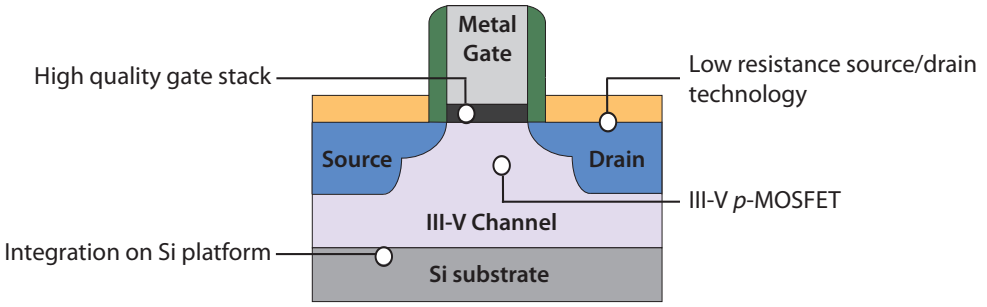


Figure 1.5: Schematic illustrating the key challenges facing III-V CMOS technology.

thermal stability. In this regard, Si is blessed with good fortune. The excellent chemical marriage between Si and its native oxide, SiO_2, results in a low interface trap density (D_{it}), on the order of $10^{10}\,cm^{-2}eV^{-1}$ [16]. Hence transistor performance is not degraded. III-Vs however are not so fortunate. The unforgiving nature of III-V native oxides results in a high density of interface defects, increased dielectric defects and increased interface roughness to the detriment of device performance. Importantly, Fermi level pinning (FLP) due to the high D_{it} will inhibit the gate's ability to modulate the channel charge [17].

An unpinned Fermi level in III-V was first reported based on the in-situ deposition of gallium-gadolinium-oxide dielectric on GaAs [18–20]. Subsequently, an ex-situ Al_2O_3 atomic layer deposition (ALD) process was demonstrated to unpin the Fermi level on GaAs [21, 22]. It was found that during the very early stages of ALD, a "self-cleaning effect" occurs that largely removes the native oxides prior to dielectric deposition [21, 22]. Since the up-take of ALD in Si CMOS manufacturing [10], ALD integration has been extensively investigated on a variety of III-V materials, most notably InGaAs [23–28] with promising results. Complementary with this, several surface passivation techniques have been examined to improve the gate stack quality including pre-deposition treatments [23–27, 29, 30], interfacial control layers [31–33], deposition chemistry [34], post-deposition treatments [23–25, 30, 35] and alternative high-k dielectrics [32, 35, 35–41] . These various investigations have led to a few demonstrations of $In_{\geq 0.53}GaAs$ n-MOSFETs with performance approaching and exceeding state-of-the-art Si n-MOSFETs [32, 42–46]. Further improvements in the gate stack are still needed before III-Vs can become a viable integration option in CMOS.

Low-resistance S/D Technology

Parasitic source/drain series resistance (R_{sd}) poses a key challenge for III-V CMOS and its future scalability. As dimensions are scaled, the channel resistance of MOSFETs decreases significantly, resulting in the on-current being eventually limited by R_{sd}. The reduction

in S/D geometries with device scaling also causes an increase in the contact resistance due to current crowding effects [47].

Two components of R_{sd} that need to be engineered to minimise the impact of parasitic resistance on device performance are the junction and contact technologies. A low damage, abrupt, conformal, heavily doped junction is required to maintain electrostatic charge control. In Si technology, ion implantation is a standard practice for junction doping [48]. This technique however poses problems for III-Vs. The alteration of the stoichiometry presents difficulties for recovering crystal damage induced by implantation. Such damage results in lower dopant activation as well as higher junction leakage [49]. A number of alternatives including selective epitaxial regrowth (SEG), monolayer doping (MLD) and plasma doping are being pursued as replacements for ion implantation [49].

In SEG, heavily doped S/D regions are regrown using either molecular beam epitaxy (MBE) or metal-organic chemical vapour deposition (MOCVD). Regrowth provides for abrupt doping profiles, high doping densities ($5.5\times10^{19}\,\mathrm{cm}^{-3}$ [50]) and reduced leakage [49]. Another promising approach is MLD. This technique comprises the termination of the III-V surface with a monolayer of dopant atoms, before dielectric capping and thermal annealing for dopant drive-in and activation [49]. MLD is a low-cost, low damage process that achieves ultra-shallow junctions of high uniformity. A third technology is plasma doping, which offers the process simplicity of ion implantation and yet provides for low damage, ultra-shallow junctions [49].

When it comes to contact technologies, III-Vs have traditionally relied on Au-based alloyed ohmic contacts. There are two drawbacks to such contacts in CMOS, the first is that Au introduces deep level traps in Si devices and therefore is barred from Si processing lines [49]. The other drawback is non-uniformity due to metal spiking in Au-based metal alloys. This makes the contact scheme incompatible for tight device geometries required in CMOS [49]. Two contact technologies have been explored on III-Vs to obviate Au-based alloyed contact schemes. The first is a process similar to silicide that is used for Si CMOS. Silicide-like contacts to III-Vs have been investigated using Ni [51–53], Co [54] and Pd [55] at relatively low temperatures. The more promising approach, yielding superior contact performance, is based on refractory ohmic contacts, e.g. Mo, to highly doped III-V materials [56–58]. It is further vital that SA techniques are used during junction/contact formation to minimise parasitic resistance [13, 14].

Integration on Si Platform

III-Vs eventually need to be heterointegrated on a Si platform. From a manufacturing perspective, a wafer-scale integration solution is highly desirable. Access to existing

mature Si technology toolsets would reduce production and substrate costs. Various integration technologies are currently under consideration. These include buffer layer growth [32, 59, 60], aspect ratio trapping (ART) [61–63] and layer transfer [64–66].

The prospect of directly growing III-Vs on Si is complicated by the large lattice mismatch between the two materials, which results in a defective grown layer. This problem can be eased by employing a buffer layer that is compositionally graded between the lattice constants of Si and III-V. Defect-free III-V layers can therefore be grown as the misfit dislocations are confined to the graded buffer. III-Vs have been successfully integrated on Si employing buffer layers ($>1\,\mu$m in thickness) grown by MBE [32, 59] and MOCVD [60] techniques. Such thick buffers are however not economical, as throughput is reduced from the long growth times, and thus not a viable manufacturing solution.

Aspect ratio trapping is another integration technique that enables III-Vs to be directly grown on Si [61–63]. In this approach, III-Vs are grown in high aspect ratio trenches defined in a SiO_2 hard mask on top of the Si substrate. The defects that nucleate at the bottom of the trench during the growth of III-Vs are trapped in the trench sidewalls, leaving the material at the top of trench, in theory, defect free. Other than demonstrations of InP and InGaAs growths, the approach has yet to yield high performance devices [61–63].

Layer transfer is the final approach for III-V integration on Si. In this method, the III-V active layer is transferred from a III-V donor wafer onto a Si acceptor wafer covered with a thin dielectric. This can be achieved by means of epitaxial layer transfer [66, 67], or by direct wafer bonding (DWB) followed by an etch back process [64] or hydrogen induced thermal splitting [65]. Czornomaz *et al.* showed that the layer transfer process can be made economically viable by recycling the donor wafer [65]. Despite this, scaling up the wafers for manufacturing purposes still poses a challenge.

III-V p-MOSFET

For CMOS logic functionality, comparably performing n- and p-MOSFETs are required. Historically, owing to the significantly higher mobility of electrons compared to holes in Si, the n-MOSFET has outperformed its counterpart. As a result, circuit designers have adapted to working with Si p-MOSFETs with only a third of the performance of the n-MOSFETs. As such, the performance gap between n- and p-devices should be no greater than this in future technology nodes.

The search for a p-channel option to complement the III-V n-MOSFET thrust reveals Ge to be an ideal candidate due to its exceptionally high hole mobility (μ_{h}), highest among all semiconductors [68], and the progress made in Ge p-MOSFET technology [69–72]. The

cointegration of Ge and III-Vs, however, significantly increases the processing complexity due to the material systems having differing chemistries and thermodynamic stabilities. Unless these issues can be addressed, Ge is not a viable p-channel option in future nodes.

An all III-V CMOS technology would greatly simplify device processing. The difficulty is finding a III-V that would satisfy the performance gap criteria. This is largely due to the significantly lower mobilities of holes compared to electrons in III-Vs [68]. There are two ways the hole mobility can be enhanced, the first being the introduction of compressive strain by means of pseudomorphic growth on materials with smaller lattice constants. This has yielded hole mobilities as high as $390\,\mathrm{cm}^2/\mathrm{Vs}$ in $\mathrm{In}_{x\geq 0.53}\mathrm{GaAs}$ [73], $1500\,\mathrm{cm}^2/\mathrm{Vs}$ in GaSb [74, 75] and InGaSb [76], and just over $1200\,\mathrm{cm}^2/\mathrm{Vs}$ in InSb [77]. Uniaxial compressive strain, similar to that used in Si CMOS to boost mobility, as well as the superposition of both uniaxial and biaxial strains have also been explored for III-Vs [78–80]. A second approach is to use a buried channel structure in which a barrier material is interposed between the channel layer and the high-k dielectric to mitigate dielectric- and interface-related scattering that causes mobility degradation [81].

As compared to the other challenges of III-V CMOS, research effort towards realising a III-V p-channel solution has been extremely limited. Despite III-V p-MOSFETs having been demonstrated in GaAs [82], InGaAs [83], GaSb [84–90] and InGaSb [67, 89, 91–93], their performance is still considerably lagging behind that of III-V n-MOSFETs. As a result, a viable III-V p-channel option is still elusive to date. This particular challenge will be the focus of this thesis.

1.3 Scope and Outline of this Thesis

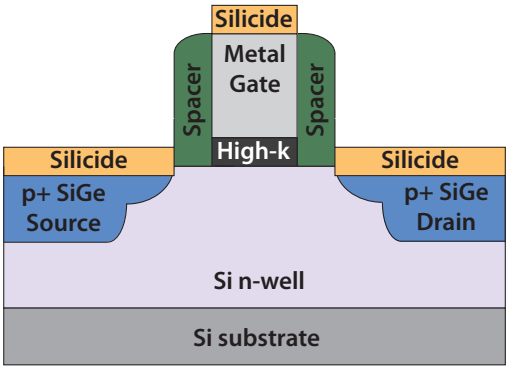
The objective of this Thesis is to develop planar III-V p-MOSFETs to determine if III-V materials are a viable p-channel solution to enable an all III-V approach for post-Si CMOS. When the project began at the end of 2008, planar device technology was expected to continue as the mainstream in CMOS. This forms the motivation to undertake planar device investigation in this work. Two essential stages of transistor development were addressed. One involves an investigation into optimising the channel transport properties by means of heterostructure engineering. The other is process module development and integration into a robust fabrication flow. This entails the development of source/drain and gate stack technologies, two key components required of any MOSFET, for the realisation of a complete device. In addition to this, the mitigation of parasitic effects on

device performance was explored via transistor layout optimisation. To this end, the specific aims of this work included:

- investigation of hole transport enhancement in III-V channels by means of layer design optimisation studies using suitably designed structures.
- development and experimental validation of source/drain ohmic contact modules appropriate for device integration.
- development and experimental validation of gate stack modules suitable for integration with III-V materials.
- investigation of methods to minimise device level parasitics which contribute to the degradation of MOSFET performance.
- demonstration of III-V p-MOSFETs based on III-V channels with the most promising hole transport and integration of developed process modules.

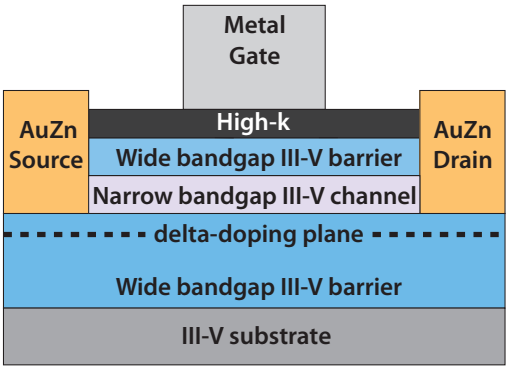
From a technological perspective, it is important to understand how the devices developed in this work are placed in the context of Si CMOS. The technologies used for Si p-MOSFETs in CMOS and for the realisation of III-V p-MOSFETs of this work are compared in Figure 1.6. Apart from the high-k/metal gate (HKMG), it is observed that the III-V devices differ from Si devices in all other technological fronts. Unlike the inversion-type Si transistors, the III-V p-MOSFETs are based on the flatband device architecture; the structure consists of a modulation-doped, high mobility channel, clad by wider bandgap barrier layers [94]. The flatband architecture obviates the need for ion-implanted source/drain (S/D) regions. This is because in the access regions on either side of the gate, which act like shallow S/D extensions in Si MOSFETs, the carriers are supplied to the channel from the ohmic contacts and the doping plane [94]. The doping arrangement further provides for low-resistance access regions [94]. As opposed to local strain in Si p-MOSFETs, achieved by recess etching and selective epitaxial regrowth of SiGe in the S/D regions to induce uniaxial compressive stress to the Si channel [9], the III-V p-channel transistors are subject to global strain. This is afforded by the flatband device architecture in which biaxial compressively stressed channels are realised through pseudomorphic growth on III-V materials with a smaller lattice constant, similar to HEMT technology. Other deviations from Si CMOS technology include growth of the III-V epi-layers on a III-V substrate rather than a Si substrate, and the adoption of lithographically-aligned (LA) AuZn contact metallisation in the S/D regions. The problem with Au-based contact technology is that it is forbidden in Si device processing, since Au acts as an impurity, producing deep-level traps in Si, thereby degrading the device performance [49]. In Si

(a) Si *p*-MOSFET (CMOS)



- Inversion-type device
- High-k/metal gate (HKMG)
- Local strain (SiGe epitaxial regrowth)
- ion-implantation for source/drain (S/D)
- Self-aligned (SA) silicide (salicide) S/D contacts
- Si substrate

(b) III-V *p*-MOSFET (this work)



- Flatband-mode device
- High-k/metal gate (HKMG)
- Global strain (pseudomorphic growth)
- no implantation for source/drain (S/D)
- Lithographically-aligned (LA) AuZn S/D contacts
- III-V substrate

Figure 1.6: Technologies used for (a) Si *p*-MOSFETs in CMOS and (b) III-V *p*-MOSFETs developed in this work.

CMOS, a self-aligned silicide (salicide) contact metallisation is used whereby a blanket deposited metal selectively reacts with the underlying semiconductor material to form a highly metallic region [48]. Metal such as Ni, Co or Ti are used in a two-step annealing process to prevent the spacer overgrowth of the silicide which could cause a leakage path between the gate and S/D regions [48]. In the first step, a low temperature anneal is performed to achieve a silicide phase of high resistivity. Following the wet chemical etching of the unreacted metal, a second higher temperature anneal is performed to realise a low resistivity silicide phase [48]. Similar salicide-like contact metallisations have also been investigated on III-V devices [51–55]. It is worth emphasising that III-V integration on a Si substrate and the realisation of a low-resistance contact technology that avoids Au, key challenges that are currently being addressed to enable a manufacturable III-V CMOS technology to be realised (see Section 1.2), did not fall into the scope of this work and thus not addressed. The objective of this work was to evaluate the viability of III-Vs as a *p*-channel solution for post-Si CMOS via device demonstrations. This required the development of process modules as relevant for device integration, without the stringent requirement of Si-compatible processing in the first instance. However, where possible Si-compatible processing was employed. Subsequent to the work presented in this Thesis,

the author has realised III-V p-MOSFETs based on a Ni-based (Au-free) S/D metallisation, similar to silicide-like contact technology, thus demonstrating the potential of the devices to be compatible with Si CMOS processing.

This work was supported by customisation funding from the Semiconductor Research Corporation (SRC) in USA, to complement the III-V n-channel work undertaken by the SRC Nonclassical CMOS Research Center (NCRC) comprising University of California Santa Barbara, University of California San Diego, Stanford University and University of Glasgow.

The organisation and contents of this Thesis are as follows: Chapter 2 provides the theoretical background needed for the understanding and analysis of devices described in this Thesis. The chapter begins with the introduction of MOSFET theory followed by a description of the metal-oxide-semiconductor (MOS) capacitor, the device channel and metal-semiconductor contacts. Chapter 3 provides a description of the fabrication techniques used for process module development and device integration which includes lithography, pattern transfer, and thin film deposition and removal. Chapter 4 presents the various characterisation methods used in evaluating the metrics related to the grown epitaxial material, ohmic contacts, gate stack and the overall device. Specifically the Hall effect and van der Pauw techniques for material transport characterisation, transmission line model for contact characterisation and MOS device characterisation for extraction of key figures of merit are described. Chapter 5 outlines the investigation of bandgap and strain engineering, collectively heterostructure engineering, for the optimisation of hole transport properties in III-V channels, using suitably designed structures. Chapter 6 summarises the process modules that were developed and examined for S/D and gate stack engineering as required for device integration. Chapter 7 describes experimental results of the fabricated III-V p-channel MOSFETs and provides an in-depth analysis of the performance metrics. Chapter 8 summarises the outcomes and key findings of this thesis and provides suggestions for future work.

2

Background Theory

2.1 Introduction

This chapter gives an overview of the key fundamentals and concepts required for the understanding and analysis of devices described in this work. The chapter begins by introducing the basic operation and current-voltage characteristics of MOSFETs, followed by a summary of short-channel effects. The key components of a MOSFET are then considered separately in three parts. In the first part, the electrostatics and capacitance-voltage characteristics of an ideal MOS capacitor are described prior to introducing the non-ideal effects. The second part deals with the carrier transport in the device channel. An overview of scattering mechanisms followed by methods to engineer the channel for improved transport are presented. Finally the chapter concludes with a section on the theory of metal-semiconductor contacts.

2.2 The Metal-Oxide-Semiconductor Field-Effect-Transistor

2.2.1 Basic Operation

A traditional bulk p-MOSFET is shown in Figure 2.1. The transistor comprises two heavily p-doped semiconductor regions, referred to as source and drain, which are separated by a n-doped semiconductor region called the substrate. This describes a p-channel MOSFET. For a n-channel MOSFET, the doping in the S/D and substrate regions would be reversed. The region between the source and drain regions, referred to as the channel, is covered by an insulator layer. A metal electrode called the gate sits atop the insulator. The insulator, referred to as the gate oxide or dielectric, isolates the channel from

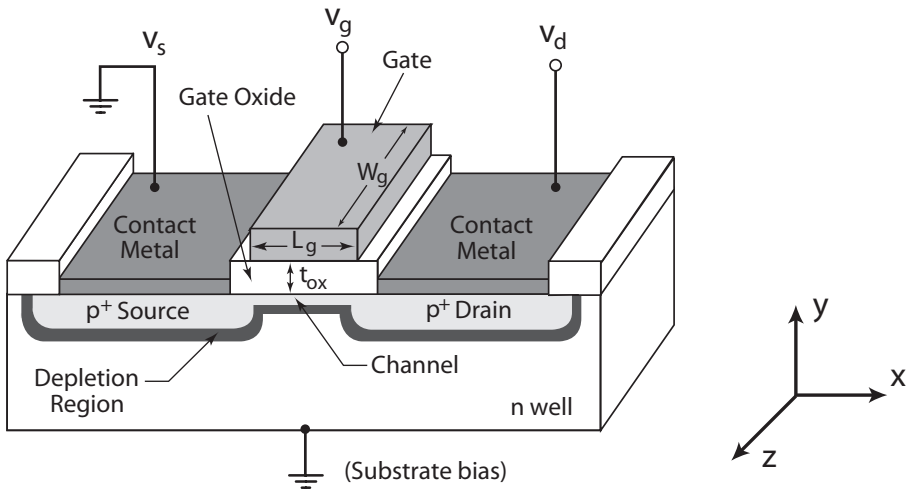


Figure 2.1: Schematic diagram of a p-channel MOSFET.

the gate. A simple MOSFET is a three-terminal device comprising a gate, source and drain contacts, each of which can be biased independently (assuming the substrate is grounded). Here, transistor operation is described for a p-channel MOSFET. Typically, the source is grounded (V_{s}=0) and the drain is biased with a negative voltage, V_{d}. Biasing the gate at zero volts causes majority carries (electrons) to accumulate underneath the gate dielectric, thereby isolating the source and drain. As a result, there is no drain current (I_{d}) flowing along the channel and the transistor is turned off. This is known as the off-state condition. As the gate voltage (V_{g}) is made more negative, electrons are repelled from the dielectric/semiconductor interface. This results in the formation of a depletion layer, devoid of carriers. At a large enough negative V_{g}, holes spill out of the S/D regions to accumulate under the gate dielectric giving rise to an inversion layer. This results in a current flow in the channel between the source and drain and the transistor is turned on, a condition referred to as on-state. The gate bias at which the device turns on is called the threshold voltage (V_{th}). Further increase in gate bias beyond V_{th} acts to increase the carrier population in the conductive channel, resulting in a higher drain current. Therefore, the current that flows from the source to drain is modulated by the gate. It is this modulation of the drain current that provides for switching operation in digital logic circuits.

2.2.2 Current-Voltage Characteristics

The behaviour of an ideal long-channel MOSFET will now be described by invoking the gradual channel approximation [95]. In this approximation, the longitudinal electric field

(E_x) along the channel is taken to be is much smaller than the transverse electric field (E_y), which in turn allows for the use of the one-dimensional form of Poisson's equation. To further simplify the analysis, the minority carriers that form the inversion layer are assumed to be located as a sheet at the dielectric/semiconductor interface (charge sheet approximation). This implies no potential drop or band bending across the inversion layer. Under these approximations, the inversion charge density (Q_{inv}) in the channel, at a position x from the source, can be expressed as:

$$Q_{inv} = C_{ox} \left(V_g - V_{th} - V_c(x) \right) , \tag{2.1}$$

where C_{ox} is the dielectric capacitance, V_g is the gate bias, V_{th} is the threshold voltage and $V_c(x)$ is the potential at a distance x along the channel with respect to the source.

The average carrier drift velocity (υ_d) at a point x along the channel is related to E_x by the carrier mobility (μ), which is assumed to be a constant:

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

The product of the carrier drift velocity, inversion charge and the channel width (W_g) gives form to the drain current that flow in the channel from drain to source:

$$I_d = \mu C_{ox} W_g \frac{dV_c}{dx} \left(V_g - V_{th} - V_c(x) \right) , \tag{2.3}$$

where I_d is assumed to be constant along the channel. Integrating Equation (2.3) along the channel under the appropriate boundary conditions of $V_c(0)=0$ at the source and $V_c(L_g)=V_d$ at the drain enables an expression for the drain current to be obtained as follows:

$$I_d = \mu C_{ox} \frac{W_g}{L_g} \left[\left(V_g - V_{th} \right) V_d - \frac{V_d^2}{2} \right] , \tag{2.4}$$

where L_g is the gate length. Equation (2.4) captures the basic current-voltage (I–V) characteristic of a long-channel MOSFET for the condition $V_g \leq V_{th}$. If $0 > V_d > V_g - V_{th}$, the quadratic term in Equation (2.4) can be neglected and the drain current increases linearly with gate bias. This is the linear region for which the drain current is expressed as:

$$I_d = \mu C_{ox} \frac{W_g}{L_g} \left[\left(V_g - V_{th} \right) V_d \right] . \tag{2.5}$$

When $V_d < V_g - V_{th} < 0$, the channel becomes pinched off. Under this condition, a segment of the channel towards the drain end contains no inversion charge and is therefore

highly resistive. Further increase in drain voltage (V_{d}) does not increase the drain current. Instead, the pinch-off point extends towards the source, with additional potential being dropped across this increasing resistance. The MOSFET is said to be in saturation and the drain current becomes independent of drain voltage as stated in Equation (2.6):

$$I_{\mathrm{d,sat}} = \mu C_{\mathrm{ox}} \frac{W_{\mathrm{g}}}{2L_{\mathrm{g}}} (V_{\mathrm{g}} - V_{\mathrm{th}})^2 \ . \tag{2.6}$$

The simple models used in deriving the long-channel behaviour of MOSFETs become invalid when transistors are scaled to short gate lengths. This is due to the short-channel effects summarised in Section 2.2.3. The drain current of short-channel devices in saturation is given by:

$$I_{\mathrm{d,sat}} = W_{\mathrm{g}} C_{\mathrm{ox}} \upsilon_{\mathrm{sat}} (V_{\mathrm{g}} - V_{\mathrm{th}}) \ , \tag{2.7}$$

where υ_{sat} is the saturation velocity.

The I–V characteristics of transistors are usually obtained and plotted in two ways. Measuring the drain current as a function of drain bias at a constant gate bias gives form to the output characteristics illustrated in Figure 2.2(a). The transfer characteristics (Figure 2.2(b)), on the other hand, plots the drain current as a function of gate bias for a constant drain bias. A number of device performance metrics can be obtained from the output and transfer characteristics as detailed below:

- **Saturation drain current:** The saturation drain curent ($I_{\mathrm{d,sat}}$) is the drive current at high V_{d} and high V_{g}, normalised to the width of the gate (W_{g}).
- **Output conductance:** The output conductance (g_{d}) is a measure of drain current

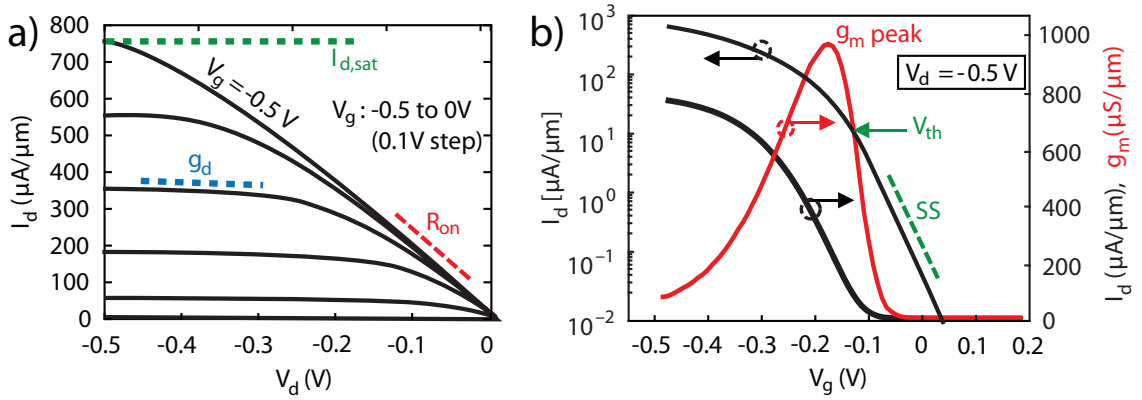


Figure 2.2: (a) Output and (b) transfer characteristics of a generic MOSFET.

saturation and is derived from the slope of the output characteristics in the saturation region ($dV_\mathrm{d}/dI_\mathrm{d}$) at medium to high drain bias.

- **On-resistance:** The drain current, in the linear region of device operation, is limited by the on-resistance (R_on), which is determined from the inverse slope of the output characteristics ($dV_\mathrm{d}/dI_\mathrm{d}$) at low V_d as shown in Figure 2.2(a). The on-resistance is comprised of the source (R_s), drain (R_d) and channel (R_ch) resistances.
- **Transconductance:** The transconductance (g_m) of a transistor is a measure of the gate control over the drive current modulation in the channel. It is defined as the rate of change of drain current with applied gate bias at a fixed drain bias:

$$g_\mathrm{m} = \frac{dI_\mathrm{d}}{dV_\mathrm{g}}\,. \tag{2.8}$$

  As indicated in Figure 2.2(b), g_m is represented by its maximum value.
- **Threshold voltage:** The threshold voltage (V_th) is defined as the gate voltage that marks the transition between the on-state and off-state operation of a transistor. The appropriate methodology for extracting V_th from the transfer characteristics will be covered in Section 4.4.2.1.
- **Subthreshold swing:** The SS, measured below V_th, quantifies the gate voltage required to switch a transistor from the off-state to the on-state. SS is defined as the gate bias required to increase the drain current by one order of magnitude, measured in mV/dec, and is determined from the inverse slope of the transfer characteristic plotted on a logarithmic scale:

$$SS = \frac{dV_\mathrm{g}}{d(log_{10}I_\mathrm{d})}\,. \tag{2.9}$$

  The theoretical limit of SS is 60 mV/dec at room temperature [5, 6].

2.2.3 Short-Channel Effects

As the channel length of a MOSFET is reduced, the spatial extent of the depletion widths of the source and drain regions becomes increasingly comparable to the channel length. This causes the source and drain regions to exact a stronger influence on the channel regions, thereby diminishing the gate control over the channel. This gives form to short-channel effects (SCE), which are discussed in the following paragraphs.

Threshold Voltage Roll-off

In a MOSFET, depletion regions are formed under the gate as well as the source and drain regions as shown in Figure 2.3(a). Based on the charge sharing model [96], fixed charges in the depletion region are balanced by equal and opposite charges both on the gate and within the source and drain depletion regions. The division and charge sharing of the depletion regions between the gate and depletion regions of the source and drain are indicated in Figure 2.3(a) for long and short-channel devices. In long-channel devices, the depletion charge is almost entirely balanced by charges in the gate, with negligible contribution from source and drain depletion regions. With decreasing channel length the source and drain depletion regions become comparable to the channel length. Therefore in short-channel devices, the source and drain depletion regions make up a larger proportion of the overall depletion region under the gate. This translates to a greater proportion of bulk depletion charge being balanced by charges within the source and drain depletion regions. As a result, the onset of inversion is achieved under a reduced charge on the gate, and therefore a lower threshold voltage, V_{th}. This reduction in threshold voltage with decreasing channel length, as illustrated in Figure 2.3(b), is referred to as V_{th} roll-off.

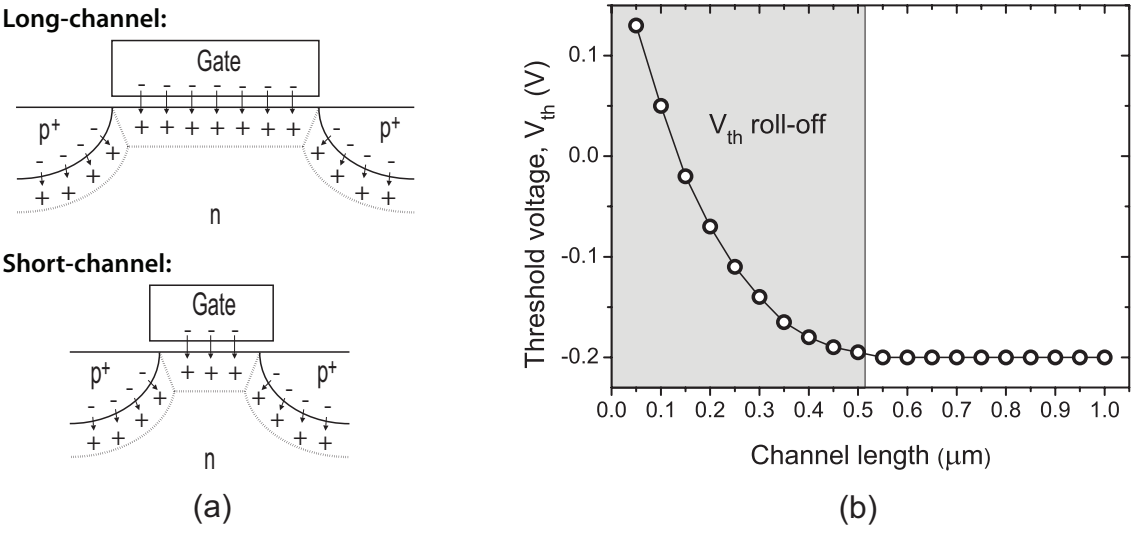


Figure 2.3: (a) Schematic illustrating charge sharing of the depletion regions in long and short-channel p-MOSFETs and (b) example plot of threshold voltage vs. channel length, based on arbitrary values, demonstrating V_{th} roll-off.

Drain-Induced Barrier Lowering (DIBL)

Another effect that arises from increased charge sharing between the gate and the depletion regions of the source and drain, due to channel length scaling, is drain-induced barrier lowering (DIBL). This occurs when the electric fields at the drain penetrate deep

into the channel, thereby lowering the potential barrier near the source. Shown in Figure 2.4(a) are the potential diagrams of long and short-channel MOSFETs for drain bias in the linear region ($V_{\mathrm{d,lin}}$) and saturation region ($V_{\mathrm{d,sat}}$) of device operation. In the case of the long-channel device, a change in drain bias has no impact on the potential barrier in the channel. In contrast, the potential barrier of the short-channel devices changes with drain bias, which induces a bias dependency to the threshold voltage. This effect causes a threshold voltage shift with increasing drain bias as shown in Figure 2.4(b). DIBL can therefore be defined as the shift of the threshold voltage, ΔV_{th}, per incremental change in the drain bias, ΔV_{d}, and is measured in mV/V:

$$\mathrm{DIBL} = \left| \frac{\Delta V_{\mathrm{th}}}{\Delta V_{\mathrm{d}}} \right| = \left| \frac{V_{\mathrm{th,lin}} - V_{\mathrm{th,sat}}}{V_{\mathrm{d,lin}} - V_{\mathrm{d,sat}}} \right| , \tag{2.10}$$

where $V_{\mathrm{th,lin}}$ and $V_{\mathrm{th,sat}}$ represent the threshold voltages in the linear and saturation regions of device operation, respectively.

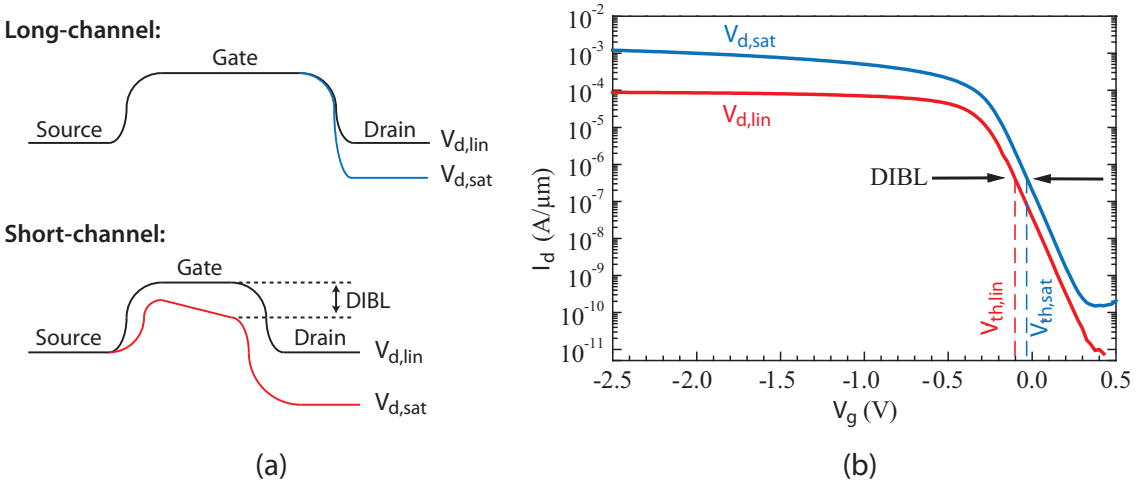


Figure 2.4: (a) Potential diagram of a long and short-channel MOSFET for drain bias in the linear region ($V_{\mathrm{d,lin}}$) and saturation region ($V_{\mathrm{d,sat}}$) of device operation and (b) example logarithmic plot of drain current vs. gate bias, based on arbitrary values, at $V_{\mathrm{d,lin}}$ and $V_{\mathrm{d,sat}}$ illustrating the DIBL effect.

Punchthrough

Punchthrough, essentially an extreme version of DIBL, occurs when the depletion region surrounding the drain extends to meet that of the source. When the two depletion regions merge, as shown in Figure 2.5(a), the space-charge-limited current flows between the source and drain. This results in the loss of gate control over channel charge modulation.

Impact ionisation

Another undesirable SCE, impact ionisation, occurs from the high velocity of carriers (hot carriers) in the presence of high longitudinal electric fields. At high fields, holes (electrons) in the valence (conduction) band acquire energy as they are accelerated towards the drain. The collision of the hot carriers with the crystal lattice results in ionisation events leading to the generation of electron-hole pairs. For high enough fields, these carriers can acquire sufficient kinetic energy to induce additional impact ionisation events. As a result, a process of carrier multiplication occurs in the high-field region near the drain. The drain current is increased from the contribution of minority carriers, while a substrate current is generated from majority carriers. The device breaks down when the longitudinal field exceeds the breakdown field of semiconductors as shown in Figure 2.5(b).

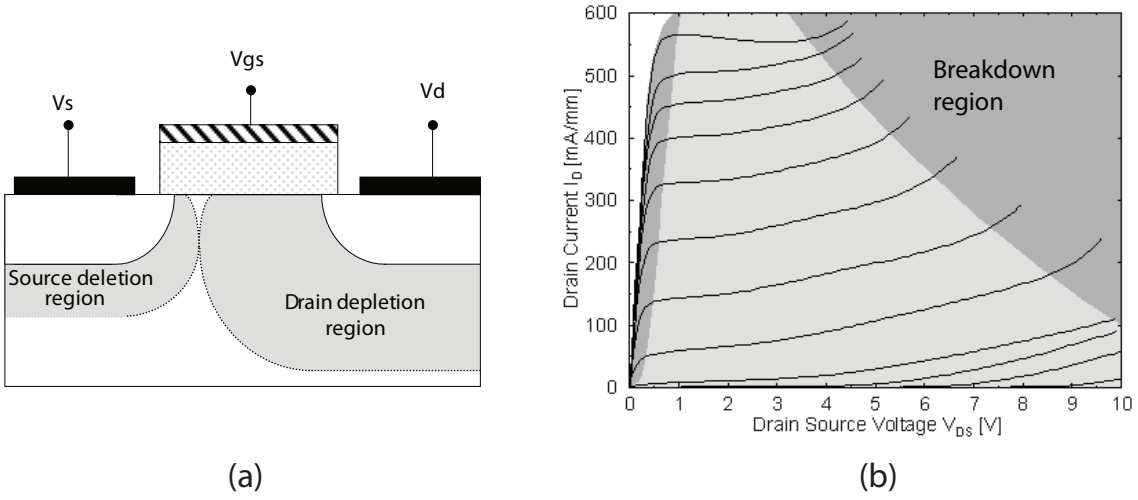


Figure 2.5: (a) Schematic diagram for punchthrough and (b) plot of drain current vs. drain voltage illustrating impact ionisation effects.

2.3 The Metal-Oxide-Semiconductor Capacitor

The metal-oxide-semiconductor (MOS) capacitor is unequivocally the core structure in digital electronics. It is used in MOSFETs to control the channel conductance by the gate bias, and therefore warrants the physics of the MOS structure to be considered. The theory presented in this section is based around an n-MOS capacitor. For a more thorough review the reader is referred to [95, 97]. It is further noted the MOS designation, typically reserved for the technologically dominant metal-SiO_2-Si system, is used throughout this thesis, for sake of simplicity, to mean structures composed of any gate dielectric or semiconductor.

2.3.1 The Ideal MOS Capacitor

The MOS capacitor is a two-terminal device comprising an oxide film sandwiched between a semiconductor substrate and a metallic field plate called the gate. Figure 2.6 shows the cross-sectional view of a MOS capacitor, where V_g is the applied gate bias. The energy

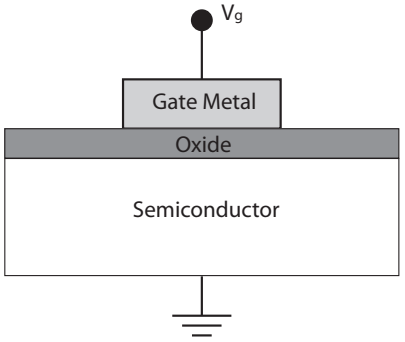


Figure 2.6: Schematic of a metal-oxide-semiconductor (MOS) capacitor.

band diagram of an ideal n-MOS capacitor, at thermal equilibrium (V_g=0), is shown in Figure 2.7 along with the definitions of notations. The MOS capacitor is called ideal if the following conditions are met:

1. For any bias condition, the only charges present in the structure are those in the semiconductor and those of equal but opposite polarity at the metal/dielectric interface.
2. The application of a bias across the MOS capacitor does not induce current flow through the gate dielectric. Only transients associated with charging/discharging the capacitor are present. Therefore, the gate dielectric is a perfect insulator.

Assuming the work function difference (ϕ_{ms}), the difference between the metal work function and the semiconductor work function, is zero, the aforementioned conditions, with the aid of Figure 2.7, translate to:

$$\phi_{ms} = \phi_m - \phi_s = \phi_m - \left(\chi_s + \frac{E_g}{2q} + \psi_B\right) = 0\,. \tag{2.11}$$

Under this condition, the energy band diagram of the MOS capacitor is flat (Figure 2.7). This implies that there is no charge on the plates of the capacitor, i.e. the metal/dielectric interface and the semiconductor, and hence there is no electric field across the oxide. This is known as the flatband condition, and the gate bias associated with this condition is referred to as the flatband voltage (V_{fb}). In the case of the ideal MOS capacitor, the

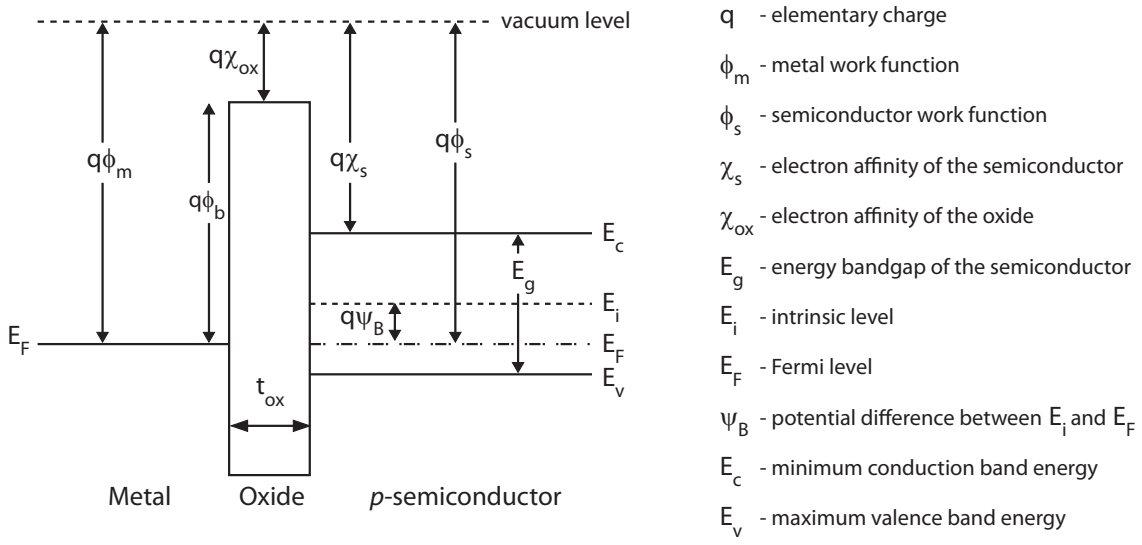


Figure 2.7: Energy band diagram of an ideal n-MOS capacitor in thermal equilibrium (V_{g}=0) and the definitions of used notations.

flatband condition is achieved under zero applied bias, thus the flatband voltage is also zero.

2.3.2 Electrostatics

The electrostatics of the ideal n-MOS capacitor for an applied bias will now be described. A detailed band diagram, illustrating the band structure and potential distribution of an ideal n-MOS capacitor under an applied bias, is shown in Figure 2.8. A potential ψ, defined as the intrinsic level ($E_{\mathrm{i}}(x)$) at any point along x and measured relative to the intrinsic level (E_{i}) in the semiconductor bulk, represents the spatial band bending in the semiconductor. In the semiconductor bulk the band bending is zero, while the surface potential (ψ_{s}), represents the band bending at the oxide/semiconductor interface. The nature and concentration of free carriers along x can be further derived from a comparison between ψ and ψ_{B}, which gives the energy position of $E_{\mathrm{i}}(x)$ relative to Fermi level (E_{F}).

When a gate bias is applied to a MOS capacitor, the potential is dropped partly across the oxide and partly across the semiconductor:

$$V_{\mathrm{g}} = V_{\mathrm{ox}} + \psi_{\mathrm{s}} \,. \tag{2.12}$$

V_{ox} is the potential across the oxide given as:

$$V_{\mathrm{ox}} = E_{\mathrm{ox}} t_{\mathrm{ox}} = \frac{|Q_{\mathrm{s}}|\, t_{\mathrm{ox}}}{\epsilon_{\mathrm{ox}}\epsilon} = \frac{|Q_{\mathrm{s}}|}{C_{\mathrm{ox}}} \,, \tag{2.13}$$

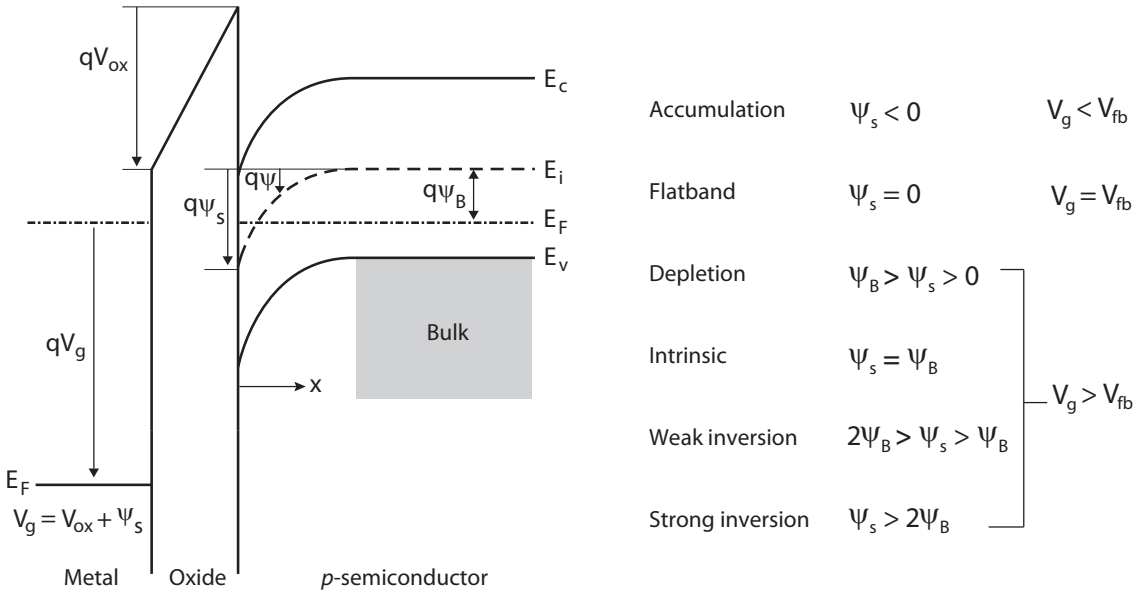


Figure 2.8: Energy band diagram of an ideal n-MOS capacitor under an applied bias: band structure, potential distribution and relationship between ψ and band bending.

where E_{ox} is the electric field across the oxide, Q_{s} is the semiconductor charge, t_{ox} is the gate dielectric thickness, ϵ_{ox} is the relative permittivity of the oxide, ϵ is the permittivity of free space and C_{ox} is the oxide capacitance. Drawing on this relationship between gate bias and surface potential, the working regions of an ideal n-MOS capacitor will now be considered.

The various working regions of an ideal n-MOS capacitor for specific biasing conditions, and hence surface potentials, are illustrated in Figure 2.8. There are namely three bias-dependent regions; accumulation, depletion and inversion, with the additional regions of flatband and intrinsic marking the transition between accumulation and depletion, and between depletion and inversion, respectively. In the following paragraphs, the bias behaviour of an ideal n-MOS capacitor is detailed with aid of the energy band diagrams and charge distributions shown in Figure 2.9.

Accumulation (ψ_{s}<0)

The application of a gate bias smaller than the flatband voltage, which in the ideal case is V_{g}<0, will deposit negative charges on the gate. To maintain charge neutrality positive charges in the semiconductor are drawn towards the oxide/semiconductor interface. The resulting accumulation of holes (majority carriers) gives rise to a change in the free carrier state as one approaches the interface and is accommodated by the energy bands in both the insulator and semiconductor bending upwards (Figure 2.9(a)). The hole concentration

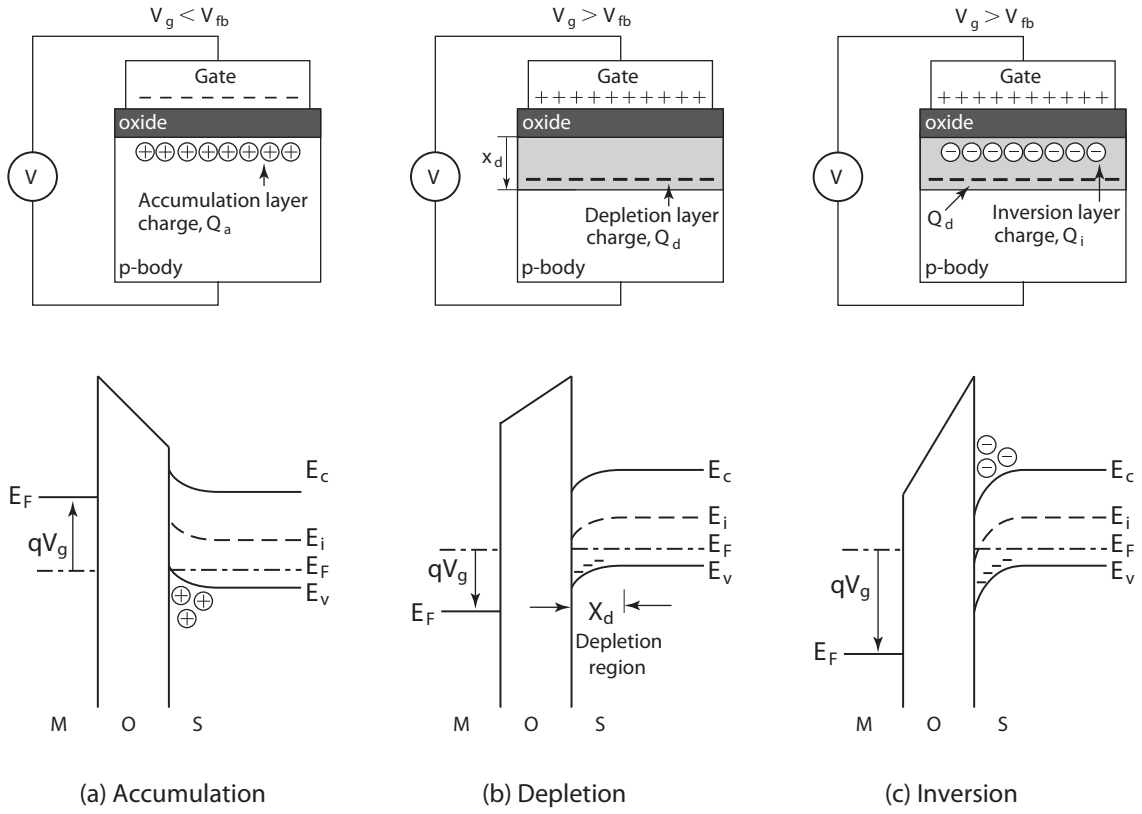


Figure 2.9: Energy band diagram and distribution of charges of an ideal n-MOS capacitor in (a) accumulation, (b) depletion and (c) inversion.

at the oxide/semiconductor interface is given by:

$$p_{\mathrm{s}} = n_{\mathrm{i}} e^{q(\psi_{\mathrm{B}}-\psi_{\mathrm{s}})/kT}, \tag{2.14}$$

where n_{i} is the intrinsic concentration, k is the Boltzmann constant and T is the absolute temperature. The intrinsic concentration is defined as:

$$n_{\mathrm{i}} = \sqrt{N_{\mathrm{c}} N_{\mathrm{v}}}\, e^{(E_{\mathrm{c}}-E_{\mathrm{v}})/2kT}, \tag{2.15}$$

where N_{c} and N_{v} represent the effective density of states in the conduction and valence band, respectively.

Depletion ($\psi_{\mathbf{B}}>\psi_{\mathbf{s}}>0$)

The case of a gate bias equal to the flatband voltage (V_{g}=0) was previously described and thus merits no further discussion. For a gate bias greater than the flatband voltage (V_{g}>0), positive charges placed on the gate are balanced by the depletion of holes from

the oxide/semiconductor interface, leaving behind negatively charged acceptor ions. The decrease of hole concentration at the interface moves the intrinsic level closer to the Fermi level, resulting in the downwards bend of the energy bands (Figure 2.9(b)). This carrier free region in the semiconductor is known as the depletion region for which the space-charge density is approximated by the density of ionised acceptors (acceptor doping concentration):

$$N_{\mathrm{A}} \;=\; n_{\mathrm{i}} e^{q\psi_{\mathrm{B}}/kT}\,. \tag{2.16}$$

Solving the one-dimensional Poisson equation within the depletion region yields the extent of the depletion into the semiconductor:

$$x_{\mathrm{d}} \;=\; \sqrt{\frac{2\epsilon_{\mathrm{s}}\epsilon\psi_{\mathrm{s}}}{qN_{\mathrm{A}}}}\,, \tag{2.17}$$

where x_{d} is the depletion layer width and ϵ_{s} is the relative permittivity of the semiconductor.

Intrinsic ($\psi_{\mathbf{s}}$=$\psi_{\mathbf{B}}$)

As the gate bias is made more positive, the energy bands will continue bending down with E_{i} getting closer to E_{F} at the oxide/semiconductor interface. At a certain bias, the intrinsic level meets the Fermi level (ψ_{s}=ψ_{B}) resulting in an intrinsic semiconductor (n_{i}) with equal hole (p_{s}) and electron (n_{s}) concentrations:

$$p_{\mathrm{s}} \;=\; n_{\mathrm{s}} \;=\; n_{\mathrm{i}}\,. \tag{2.18}$$

Inversion ($\psi_{\mathbf{s}}$>$\psi_{\mathbf{B}}$)

With a further increase in positive bias, the intrinsic level crosses the Fermi level and electrons (minority carriers) are attracted to the oxide/semiconductor interface. This causes the electron concentration to exceed the hole concentration at the interface, thereby forming an inversion layer separated from the bulk by the depletion layer (Figure 2.9(c)). The surface electron concentration is given by Equation (2.19):

$$n_{\mathrm{s}} \;=\; n_{\mathrm{i}} e^{-q(\psi_{\mathrm{B}}-\psi_{\mathrm{s}})/kT}\,. \tag{2.19}$$

Depending on the electron concentration, two bands of inversion exist; weak inversion and strong inversion. The surface is deemed to be weakly inverted when the electron concentration is greater than the intrinsic concentration but smaller than the bulk hole

concentration:

$$n_{\text{bulk}} > n_{\text{s}} > n_{\text{i}}\,, \tag{2.20}$$

where n_{bulk} is approximated as N_{A}. The onset of strong inversion occurs for the condition:

$$n_{\text{s}} = n_{\text{bulk}}\,. \tag{2.21}$$

An expression for the surface potential at the onset of strong inversion $\psi_{\text{s,sinv}}$ can then be derived by combining Equations (2.16), (2.19) and (2.21):

$$\psi_{\text{s,sinv}} = 2\psi_{\text{B}} = \frac{2kT}{q}\ln\left(\frac{N_{\text{A}}}{n_{\text{i}}}\right). \tag{2.22}$$

The depletion layer width which has been growing in size since the onset of depletion reaches its maximum limit, $x_{\text{d,max}}$, after strong inversion is reached. This means that any further increase in positive gate bias results in stronger inversion, from thermal generation of electrons, rather than in more depletion. By substituting Equation (2.22) into (2.17), the maximum depletion width is calculated as follows:

$$x_{\text{d,max}} = \sqrt{\frac{4\epsilon_{\text{s}}\epsilon kT\ln(N_{\text{A}}/n_{\text{i}})}{q^2 N_{\text{A}}}}\,. \tag{2.23}$$

2.3.3 Capacitance-Voltage Characteristics

Since direct current (DC) current flow is blocked by the oxide, the major observable exhibited by MOS capacitors is capacitance, which forms the topic of discussion here.

The capacitance-voltage (C–V) characteristics of a MOS capacitor is obtained by applying a DC gate bias, V_{g}, to the metal gate and a small amplitude ($\sim$20$-$50 mV) alternating current (AC) signal, with frequencies in the range of tens of Hz to a few MHz, is superimposed. The DC bias induces a space charge and band bending, the electrostatics of which was previously described, thus determining the Fermi level position at the interface. The AC signal induces a periodic change in band bending, causing the Fermi level at the interface to oscillate around the energy level position determined by the DC bias. The capacitance will vary depending on the applied gate bias as well as frequency.

The frequency dependence of the C–V measurement is related to the carrier response times. The majority carrier response time (τ_{maj}) is expressed as [97]:

$$\tau_{\mathrm{maj}} = \frac{q\lambda_{\mathrm{d}}^2}{\mu kT}, \tag{2.24}$$

where μ is the carrier mobility and the Debye length, λ_{d}, is given as [97]:

$$\lambda_{\mathrm{d}} = \sqrt{\frac{\epsilon_{\mathrm{s}}\epsilon kT}{q^2 N_{\mathrm{A}}}}. \tag{2.25}$$

Substituting Equation (2.25) into (2.24) gives:

$$\tau_{\mathrm{maj}} = \frac{\epsilon_{\mathrm{s}}\epsilon}{q\mu N_{\mathrm{A}}}. \tag{2.26}$$

The minority carrier response time (τ_{min}) is given by the relation [97]:

$$\tau_{\mathrm{min}} = \frac{1}{\sqrt{2}}\frac{N_{\mathrm{A}}}{n_{\mathrm{i}}}\tau_{\mathrm{L}}\sqrt{1-\frac{\psi_{\mathrm{T}}}{\psi_{\mathrm{B}}}}, \tag{2.27}$$

where $\tau_{\mathrm{L}} = \sqrt{\tau_{\mathrm{Lp}}\,\tau_{\mathrm{Ln}}}$ (τ_{Lp} and τ_{Ln} are the bulk hole and electron lifetimes respectively) is the minority carrier lifetime and ψ_{T} is the potential signifying the bulk trap level above the intrinsic level [97]. So long as the period of the AC signal is much longer than τ_{maj} ($1/\omega \gg \tau_{\mathrm{maj}}$), majority carriers will completely follow the applied signal frequency. The same applies to minority carriers.

The small-signal equivalent circuit of a MOS capacitor, shown in Figure 2.10, comprises two capacitors in series; a fixed voltage independent oxide capacitance and a voltage dependent semiconductor capacitance ($C_{\mathrm{s}}(\psi_{\mathrm{s}})$). The total capacitance (C) is then expressed as:

$$\frac{1}{C} = \frac{1}{C_{\mathrm{s}}(\psi_{\mathrm{s}})} + \frac{1}{C_{\mathrm{ox}}}. \tag{2.28}$$

The capacitance of the oxide and semiconductor can each be modelled by the parallel plate capacitor analogy, for which capacitance normalised to area is given as:

$$C = \frac{\epsilon_0\epsilon}{t}, \tag{2.29}$$

where t and ϵ_0 represent the thickness and relative permittivity of the material, respectively.

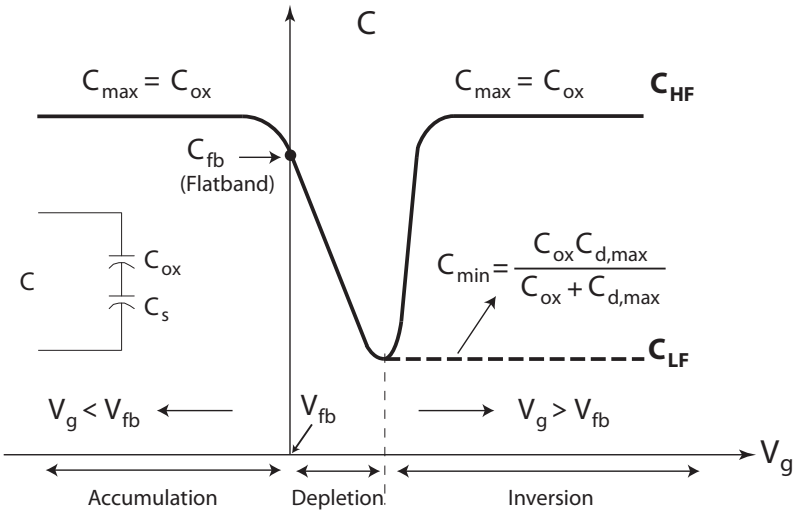


Figure 2.10: Capacitance-voltage characteristics of an ideal n-MOS capacitor.

Having established the requisite foundation, the ideal C–V characteristics of a n-MOS capacitor in accumulation, depletion and inversion is detailed below.

Accumulation

In accumulation, the only free carriers at the oxide/semiconductor interface are those of majority carriers (holes) which are in excess. Within the signal frequencies of interest, which extend up to a few MHz, $1/\omega \gg \tau_{\mathrm{maj}}$. Therefore, the majority carriers are able to completely follow the signal, resulting in a frequency independent C_{s}. As the accumulation layer is a sheet charge, the maximum capacitance (C_{max}) in accumulation is approximated by the oxide capacitance, shown in Figure 2.10, modelled as a parallel plate capacitor:

$$C_{\mathrm{max}} = C_{\mathrm{ox}} = \frac{\epsilon_{\mathrm{ox}}\epsilon}{t_{\mathrm{ox}}}. \tag{2.30}$$

Depletion

In depletion, holes (majority carriers) are repelled from the interface to form a depletion layer comprised of negative acceptor atoms. The extent to which the depletion layer proceeds into the semiconductor bulk, x_{d}, varies with gate bias (see Equation (2.17)), with the depletion width increasing with more positive bias. This gives rise to voltage dependent semiconductor capacitance in depletion, C_{d}. Based on the parallel plate analogy, the depletion layer capacitance is given by Equation (2.31):

$$C_{\mathrm{d}} = \frac{\epsilon_{\mathrm{s}}\epsilon}{x_{\mathrm{d}}}. \tag{2.31}$$

In this case, the excess of majority carriers near the depletion layer edge are able to follow the AC signal frequencies up to MHz frequencies, giving form to frequency independent C_{d}. The depletion capacitance (C_{dep}) can be modelled as the oxide capacitance in series with the depletion layer capacitance according to Equation (2.28) to give:

$$C_{\mathrm{dep}} = \frac{C_{\mathrm{ox}} C_{\mathrm{d}}}{C_{\mathrm{ox}} + C_{\mathrm{d}}} \,. \tag{2.32}$$

Due to the voltage dependence of C_{d}, the depletion capacitance decreases with increasing positive gate bias as seen in Figure 2.10.

Inversion

The C–V characteristic in inversion is dependent on the measurement frequency of the applied AC signal with respect to the thermal generation-recombination rate of minority carriers in the inversion layer, characterised by τ_{min}, and can be split into two cases:

1. For $1/\omega \gg \tau_{\mathrm{min}}$, minority carriers are able to respond completely to the AC signal. Similar to the accumulation layer, the inversion layer is a sheet charge with zero thickness. As a result, the inversion capacitance (C_{inv}) rapidly increases and saturates at a maximum value that approaches the oxide capacitance (see Equation (2.30)), just as in the accumulation case, as shown in Figure 2.10. Such an admittance response is referred to as low-frequency (LF) behaviour.
2. For $1/\omega < t_{\mathrm{min}}$, the minority carriers are no longer able to follow the signal, and thus do not contribute to the small-signal AC capacitance. This is similar to the depletion case with one difference. Unlike in depletion, where the depletion layer width changes with gate bias, in inversion the depletion width reaches a maximum value, $x_{\mathrm{d,max}}$ (see Equation (2.23)). The corresponding maximum depletion layer capacitance ($C_{\mathrm{d,max}}$) is given by:

   $$C_{\mathrm{d,max}} = \frac{\epsilon_{\mathrm{s}} \epsilon}{x_{\mathrm{d,max}}} \,. \tag{2.33}$$

   Hence C_{inv} saturates at a minimum value given by the series combination of the oxide capacitance (C_{ox}) and $C_{\mathrm{d,max}}$:

   $$C_{\mathrm{min}} = \frac{\epsilon_{\mathrm{s}} \epsilon_{\mathrm{ox}} \epsilon}{\epsilon_{\mathrm{ox}} x_{\mathrm{d,max}} + \epsilon_{\mathrm{s}} t_{\mathrm{ox}}} \,. \tag{2.34}$$

   This characteristic is illustrated in Figure 2.10 and represents the high-frequency (HF) C–V behaviour.

2.3.4 Non-Ideal Effects

So far the description of the MOS capacitor has been based on the ideal case, which provided a convenient platform for establishing the basic principles of MOS theory in a clear and simple manner. Now we depart from this to discuss the non-ideal effects observed in practice which cause a deviation from the ideal behaviour.

2.3.4.1 Charge Quantisation and Finite DOS

Earlier, the inversion layer charge was treated as a sheet charge of zero thickness at the oxide/semiconductor interface. In actual fact the peak charge density (centroid) is not located at the interface but rather at some distance inside the semiconductor, as shown in Figure 2.11(a), and is attributed to the two-dimensional quantisation of charge carriers in the inversion layer [98, 99]. The capacitance associated with the inversion layer thickness (t_{cen}) is given by Equation (2.35):

$$C_{\text{cen}} = \frac{\epsilon_{\text{s}}\epsilon}{t_{\text{cen}}}, \tag{2.35}$$

where the centroid capacitance, C_{cen}, is the capacitance due to charge quantisation. This capacitive component effectively increases the capacitive equivalent thickness (CET). CET is defined as the SiO_2 thickness (with a dielectric constant of 3.9) corresponding to the measured C_{max} in inversion or accumulation for any dielectric/semiconductor MOS capacitor and is given by [99]:

$$\text{CET} = \frac{3.9\,\epsilon}{C_{\text{max}}}. \tag{2.36}$$

The capacitance of any gate dielectric can also be expressed as an equivalent thickness of SiO_2, referred to as the equivalent oxide thickness (EOT) [99]. For a single layer of dielectric, without any interfacial layer, EOT is given as:

$$\text{EOT} = \frac{3.9\,t_{\text{ox}}}{\epsilon_{\text{ox}}}. \tag{2.37}$$

In the case of the ideal MOS capacitor C_{max} in Equation (2.36) is replaced by C_{ox}, which results in CET=EOT. In the presence of the inversion layer thickness, however, C_{max} is given by the series combination of C_{ox} and C_{cen}. This in turn gives rise to a CET value which is larger than the EOT value.

Another capacitive contribution in inversion is the density of states capacitance (C_{dos}) or quantum capacitance. This originates from the finite density of states (DOS) in the conduction/valence bands [98, 99]. Due to the finite DOS, a finite amount of change in

the surface potential, and therefore a change in bias, is required to increase the carrier concentration. The quantum capacitance is given by [100]:

$$C_{\text{dos}} = \frac{4\pi m^* q^2}{h^2}, \tag{2.38}$$

where m^* is the effective carrier mass and h is Planck's constant. Therefore, the equivalent inversion capacitance is the series combination of the oxide capacitance, centroid capacitance and quantum capacitance, as shown schematically in Figure 2.11(b):

$$\begin{aligned} \frac{1}{C_{\text{inv}}} &= \frac{1}{C_{\text{ox}}} + \frac{1}{C_{\text{cen}}} + \frac{1}{C_{\text{dos}}} \\ &= \frac{t_{\text{ox}}}{\epsilon_{\text{ox}}\epsilon} + \frac{t_{\text{cen}}}{\epsilon_{\text{s}}\epsilon} + \frac{h^2}{4\pi m^* q^2}. \end{aligned} \tag{2.39}$$

As seen from Equation (2.39), C_{max} in inversion is reduced from the value of C_{ox}, thereby invalidating the previous approximation of C_{max}=C_{ox}. As the gate oxide is scaled and higher-k oxides are adopted, the "electrical" oxide thickness ($t_{\text{ox}}/\epsilon_{\text{ox}}$) approaches and becomes comparable to the "electrical" inversion layer thickness ($t_{\text{cen}}/\epsilon_{\text{s}}$). This results in a significant portion of the capacitance in inversion being dominated by C_{cen} and C_{dos}. The effects of charge quantisation and a finite DOS will also be more prominent in semiconductors with a smaller effective mass [99]. It is further noted that the aforementioned analysis apply to the capacitance in the accumulation region.

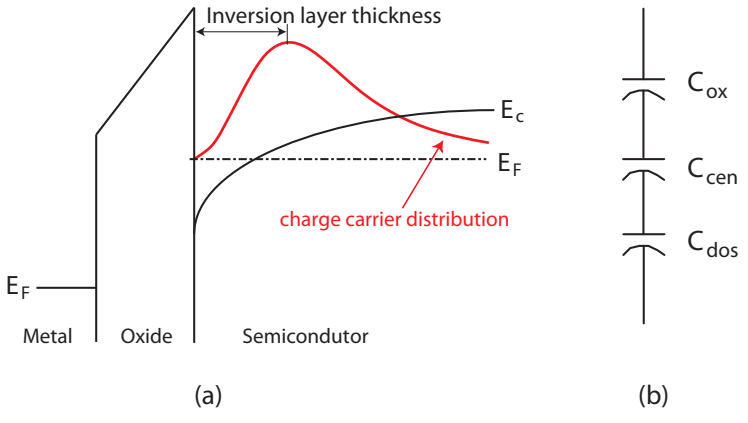


Figure 2.11: (a) Schematic illustrating the effect of charge quantisation on the location of the charge centroid in inversion and (b) small-signal equivalent circuit of the MOS capacitor in inversion.

2.3.4.2 Work Function Difference

In real MOS structures, there is a mismatch between the work functions of the metal and the semiconductor, from the choice of gate metal and the doping level in the semiconductor. For the purpose of discussion the work function mismatch will be taken to be negative (ϕ_{ms}<0). In equilibrium (V_{g}=0), due to the requirement of a common Fermi level, an electric field develops across the oxide to accommodate the work function mismatch. In doing so, the gate is positively charged and the semiconductor surface is negatively charged, thereby causing a downward band bending. This is illustrated in Figure 2.12(a). To recover the flatband condition (ψ_{s}=0), shown in Figure 2.12(b), a negative bias must be applied to the gate metal relative to the semiconductor and is given by:

$$V_{\mathrm{fb}} \;=\; \phi_{\mathrm{ms}}\,. \tag{2.40}$$

Owing to the work function difference, the flatband condition is no longer attained in equilibrium but at V_{g}<0. This causes a negative shift of the ideal C–V curve as shown in Figure 2.12(c). In the case of a positive ϕ_{ms} mismatch the ideal C–V shifts positive, resulting in the flatband condition being obtained at V_{g}>0.

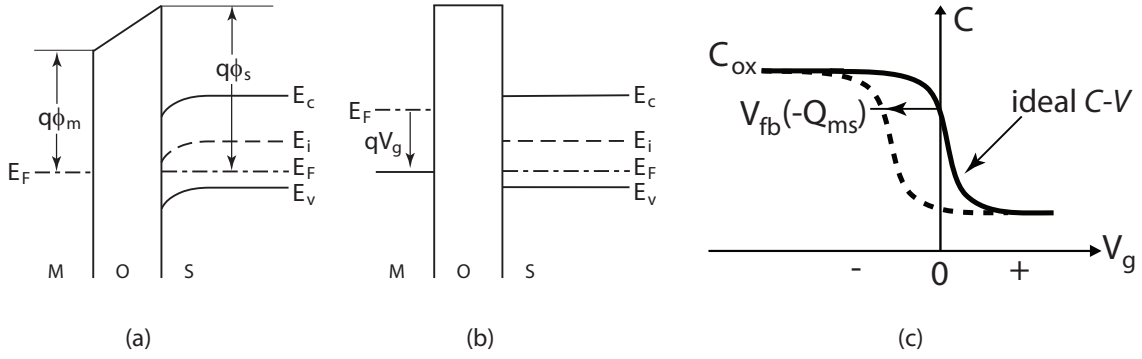


Figure 2.12: Effect of a negative work function difference (ϕ_{ms}<0) for a n-MOS capacitor: (a) band bending in equilibrium (V_{g}=0), (b) flatband condition achieved under negative gate bias (V_{g}=V_{fb}=ϕ_{ms}) and (c) negative shift of C–V curve.

2.3.4.3 Charge Defects

Experimentally, charge defects present in the oxide/semiconductor system introduce deviations to the ideal C–V characteristics and can alter device performance considerably. Various types of defects can arise in the gate stack, as shown in Figure 2.13, and can be broadly classified as *interface defects*, that reside at the oxide/semiconductor interface, and *oxide defects*, located throughout the oxide. In addition to the location, charge defects

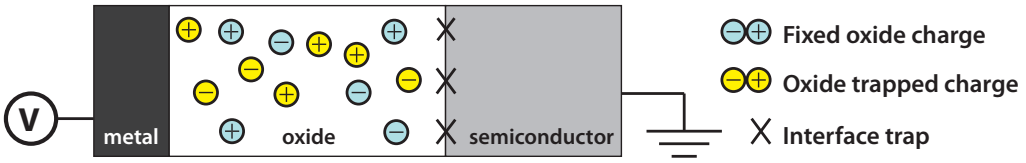


Figure 2.13: Schematic illustrating charge defects in the MOS structure.

can be further classified depending on their electrical behaviour and charge condition as discussed below.

Interface Defects

Interface defects, also known as interface traps, interface states, fast states, and surface states, are a result of structural imperfections at the surface of the semiconductor crystal lattice. These traps, distributed in energy at the oxide/semiconductor interface, can communicate with the underlying substrate over a wide range of time scales. Carrier exchange with the semiconductor occurs by a thermally activated emission/capture process [97] and the occupancy of traps is determined by Fermi-Dirac statistics [97, 101]. The associated characteristic trapping lifetime (τ_{it}) for trap occupation is given by [97, 101]:

$$\tau_{\mathrm{it}} = \frac{e^{(E_{\mathrm{maj}}-E_{\mathrm{tr}})/kT}}{\sigma \upsilon_{\mathrm{t}} N_{\mathrm{maj}}}, \tag{2.41}$$

where E_{maj} is the majority carrier band edge energy, E_{tr} is the interface trap energy position, υ_{t} is the thermal velocity of the majority carriers, σ is the capture cross section of the trapping state and N_{maj} is the density of states in the majority carrier band. The net effective number of interface trap charges per unit area per electron volt (number/cm^{2}eV) at the interface is referred to as the interface trap density (D_{it}).

Interface traps can either be donors or acceptors, and presumably both types of traps are present at every interface. Donors are neutral when occupied (0) or positively charged when empty (+). On the other hand, acceptors are neutral when empty (0) and become negatively charged when occupied (-). All traps below the Fermi level are occupied while above the Fermi level only empty traps exist. The nature and distribution of interface traps, and hence their occupancy, is dependent on the particular oxide/semiconductor system in question. Assuming donor-like traps in the lower half and acceptor-like traps in the upper half of the bandgap, the dependence of trap occupancy on the surface potential is illustrated in Figure 2.14 for a n-MOS device. At flatband, donors below E_{F} are occupied by electrons and thus neutral. The donor traps between E_{i} and E_{F} are unoccupied and thus positively charged. Above E_{F} only neutral acceptors remain. Therefore, interface

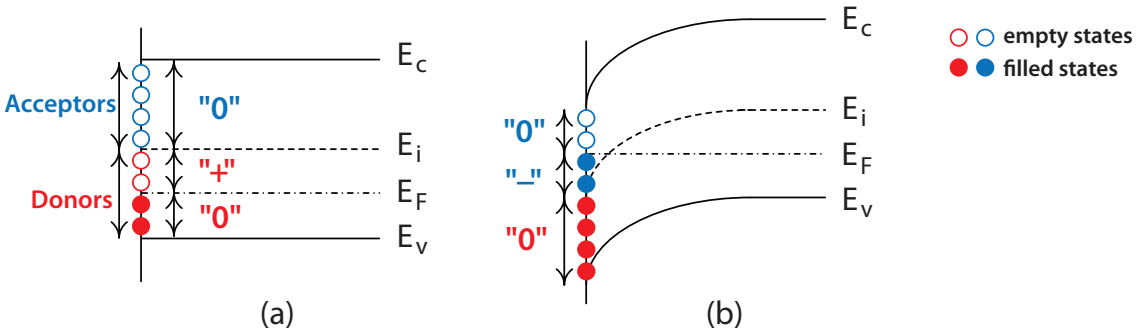


Figure 2.14: Band diagrams of a n-MOS device showing the occupancy of interface traps and the various charge polarities for a p-semiconductor with (a) positive interface trap charge at flatband and (b) negative interface trap charge at inversion. Interface traps are either occupied by electrons (solid circle) or empty (open circles).

traps contribute a net positive charge at flatband (Figure 2.14(a)). For a positive gate bias, $\psi_s > \psi_B$, all donors are neutral with a fraction of acceptors occupied by electrons. This results in a net negative charge in inversion (Figure 2.14(b)).

The occupancy of traps with changing gate bias, and hence a charge change, will distort the C–V characteristics in a few ways, depending on the frequency of measurement. To analyse these effects, the simple equivalent circuit of the ideal MOS capacitor, modelled as C_{ox} in series with C_d, can be modified to include the interface trap capacitance (C_{it}), as shown in Figure 2.15(a). At low frequencies when $\tau_{it} < 1/\omega$, all interface traps are able to respond to the AC signal frequency. The LF capacitance (C_{LF}) is thus given by:

$$\frac{1}{C_{LF}} = \frac{1}{C_{ox}} + \frac{1}{C_d + C_{it}}. \tag{2.42}$$

As seen from Equation (2.42), the additional capacitive contribution from interface traps increases the measured LF capacitance for a given gate bias as shown in Figure 2.15(c). In addition to an AC response, interface traps also follow the slowly varying DC gate bias by exchanging charge with the semiconductor every time the surface potential crosses the interface trap energy position (E_{tr}). This results in a reduced band bending under a given gate bias. Therefore, more charge, and thus a greater applied bias, is needed to obtain a similar surface potential achieved in the absence of traps. As a result, the C–V characteristic becomes stretched out in the voltage axis (Figure 2.15(c)). At high frequencies when $\tau_{it} >> 1/\omega$, the interface traps are not able to follow the AC signal and therefore, do not contribute a capacitance. As a result, the equivalent circuit reduces to the simple series approximation shown in Figure 2.15(b) and the HF capacitance (C_{HF})

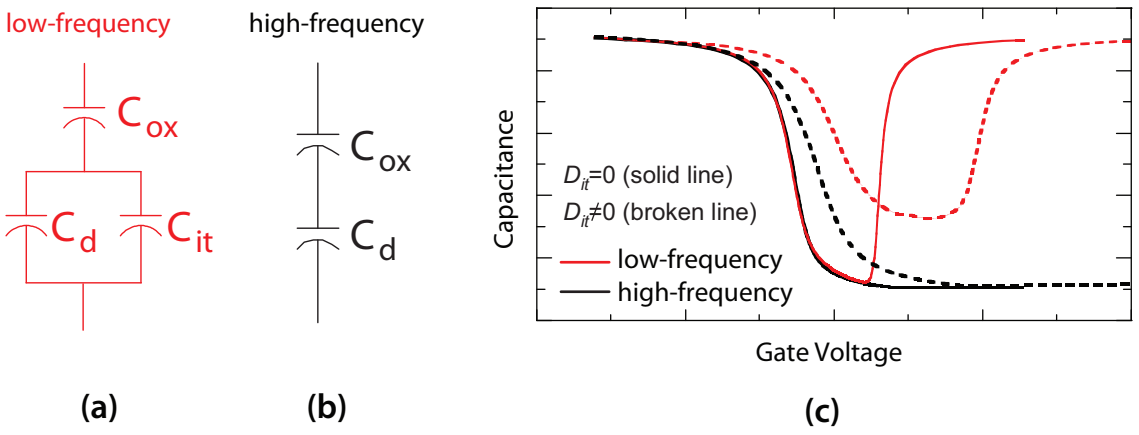


Figure 2.15: Effect of D_{it} on n-MOS capacitor: small-signal equivalent circuit at (a) low-frequency and (b) high-frequency measurements, and (c) low- and high-frequency C–V characteristics.

is given by:

$$C_{\mathrm{HF}} = \frac{C_{\mathrm{ox}} C_{\mathrm{d}}}{C_{\mathrm{ox}} + C_{\mathrm{d}}} \,. \tag{2.43}$$

Although interface traps do not contribute any AC capacitance, they still respond to the DC bias during the high-frequency measurement. To accommodate the extra charge from interface traps, similar to the LF curve, the C–V curve is stretched out along the voltage axis for HF measurements (Figure 2.15(c)).

Oxide Defects

Oxide defects include traps and charges spatially distributed in the dielectric bulk. Two types of charge defects reside in the oxide as shown in Figure 2.13; oxide trapped charge and fixed oxide charge.

Oxide trapped charges are usually neutral and become charged by trapping electrons or holes introduced into the dielectric film during fabrication and/or device operation. Oxide trapped charges can be further differentiated as *border traps* or *bulk oxide traps* based on the spatial distribution of the traps from the oxide/semiconductor interface.

Border traps, also know as near-interfacial oxide traps, slow states and switching oxide traps, are defects that are located in close proximity to the interface. While there is no distinct limit to the distance these traps can extent into the oxide, border traps are considered to be traps that can electrically communicate with the substrate through a temperature-independent tunnelling process that depends only on the distance of the defects from the interface and the measurement frequency [102, 103]. The characteristic time constants of these traps will decrease exponentially with distance from the interface,

resulting in an almost exponential decrease in the number of traps participating with measurement frequency [102]. Border traps therefore cause a frequency dispersion in the measured capacitance in accumulation (Figure 2.16(a)). They also manifests as hysteresis in the C–V response as shown in Figure 2.16(b). This is attributed to the difference in gate bias between electrons filling and emptying the traps and also to the difference in capture and emission time constants of border traps [104].

Bulk oxide traps, on the other hand, are those traps that are located at a far enough distance from the interface that at all measurement frequencies, due to the long time constants, they are not in electrical communication with the semiconductor. Consequently, the only effect of such traps is the positive or negative shift of the C–V curve along the voltage axis, depending on the net charge character, as shown in Figure 2.16(c). It is worth noting that as the gate oxide is scaled the effects of border traps become more prominent. Moreover, the distinction between border and bulk oxide traps becomes blurred since all traps distributed in scaled dielectrics will be in electrical communication with the substrate. As a result, border and bulk oxide traps are used synonymously to refer to electrically active traps.

Another type of oxide defect is fixed oxide charges, which can be positively or negatively charged. As the name suggests, such charges are unable to free themselves from the dielectric film after being trapped. Hence fixed oxide charges are not in electrical communication with the underlying substrate. Similar to bulk oxide traps, fixed charges cause the C–V curves to shift parallel, with the direction of shift determined by the charge polarity (Figure 2.16(c)).

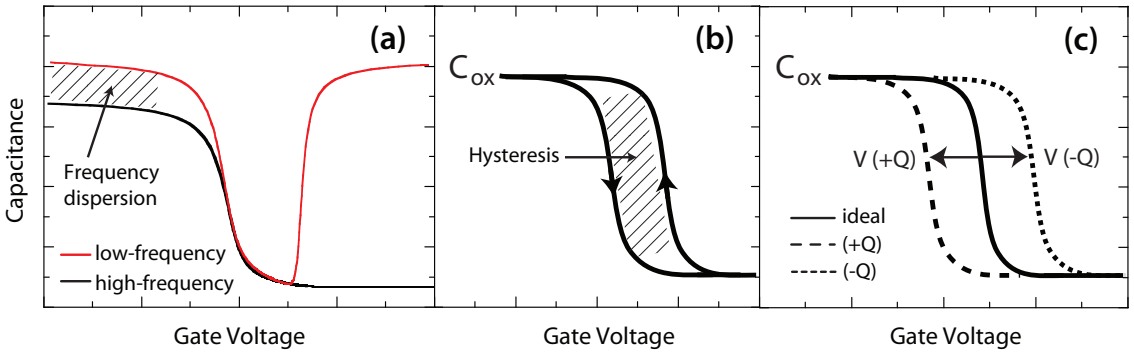


Figure 2.16: Effect of oxide defects on C–V characteristics of an n-MOS capacitor: (a) frequency dispersion in accumulation due to border traps, (b) hysteresis during the forward and reverse C–V sweeps due to border traps and (c) negative or positive C–V shift from border traps, bulk oxide traps and fixed oxide charges.

2.4 Device Channel

The performance potential of MOSFETs originates at the material level defined by the channel transport properties of carrier mobility and concentration. Arguably the mobility is the single-most important parameter with regards to device performance. As the transistor drain current is proportional to the mobility (see Equation (2.4)), the drive current would see improvement from increasing the mobility. This is particulary important in III-V p-channel devices due to the inherently poor hole mobility of III-V materials.

The carrier mobility (μ) is defined as the proportionality of the carrier drift velocity (υ_{d}) to the electric field strength (E_{x}), at low electric fields (see Equation (2.2)). The mobility can also be related to the mean free time (τ_{m}) or mean free path (λ_{m}) as follows:

$$\mu = \frac{q\tau_{\mathrm{m}}}{m^*} = \frac{q\lambda_{\mathrm{m}}}{m^*\upsilon_{\mathrm{t}}}, \tag{2.44}$$

where m^* is the effective mass and υ_{t} is the average thermal velocity of carriers given by:

$$\upsilon_{\mathrm{t}} = \sqrt{\frac{3kT}{m^*}}. \tag{2.45}$$

Equation (2.44) highlights the need to minimise carrier scattering, for increased mean free time, and reduce carrier effective mass to attain high carrier mobilities. In this section, carrier scattering mechanisms are briefly introduced before looking at techniques to engineer the channel for enhanced hole mobility.

2.4.1 Carrier Scattering - A Brief Overview

A number of scattering events can impede the motion of carriers as they traverse the channel between the source and drain of a transistor, thereby reducing the carrier mobility, of detriment to drive current performance. The main scattering events [105, 106] are briefly highlighted below:

- **Ionised Impurity Scattering:** The presence of ionised dopants in the semiconductor gives rise to ionised impurity scattering. Scattering occurs from Coulombic interactions of carriers with the electric fields of ionised dopants. The probability of the scattering event increases at higher carrier densities, thus becoming a dominant mobility-limiting scattering mechanism. The scattering is also more effective at reduced temperatures. This stems from the drop in the thermal velocity of carriers with a decrease in temperature (see Equation (2.45)), thereby causing the carriers to

spend a longer time in the vicinity of the ionised dopants and thus more susceptible to Coulomb scattering.

- **Neutral impurity scattering:** The carrier interactions with unionised impurities gives form to neutral impurity scattering. This scattering event generally has a weak contribution.
- **Alloy scattering:** Alloy scattering takes place in ternary semiconductors. The inherent aperiodicity, due to the random distribution of the alloy components in the crystal lattice, gives rise to carrier scattering.
- **Interface scattering:** The non-ideal effects of roughness and defect states at the interface between two materials result in interface scattering. This scattering mechanism is an important consideration in thin films and MOS systems.
- **Optical phonon scattering:** The vibration of the crystal lattice at finite temperatures is associated with the relative displacement of atoms within a unit cell caused by the deformation potential - phonon scattering. The high momentum of the optical phonons can have a severe influence on channel transport.
- **Carrier-carrier scattering:** This scattering mechanism takes place at high carrier densities. As the total momentum of the carrier gas is unchanged, the scattering event has little impact on the mobility. However, in conjunction with other enhanced scattering mechanism, carrier-carrier scattering can be influential.

2.4.2 Channel Engineering

2.4.2.1 Modulation-Doped Heterostructure

The ability to realise heterostructures, by means of epitaxial growth techniques such as MBE and MOCVD, has provided for improved channel transport in devices. An example of such a heterostructure-based device is the high hole mobility transistor (HHMT) [107]. The underlying principle behind the channel conduction in HHMTs is the creation of a two-dimensional hole gas (2DHG). This arises from the heterojunction formation between two dissimilar semiconductor materials, for example a smaller bandgap, undoped $In_{0.53}Ga_{0.47}As$ and a larger bandgap, p-doped $In_{0.52}Al_{0.48}As$. When the two isolated materials are brought into intimate contact, alignment of the Fermi level occurs. Due to the discontinuity in the valence band, the higher energy holes in the p-doped $In_{0.52}Al_{0.48}As$ will diffuse across the interface and collect close to the interface in the $In_{0.53}Ga_{0.47}As$. The dipole between the negatively charged acceptor ions and the accumulated holes results

in an electric field. Equilibrium is established as shown in Figure 2.17 when this electric field opposes further hole diffusion. A strong electric field caused by a large accumulation of holes results in strong band bending such that the valence band edge close to the $In_{0.53}Ga_{0.47}As$ interface is raised above the Fermi level. This in turn leads to the holes being confined in a narrow potential well occupying discrete subband levels as illustrated in Figure 2.17. Due to quantum confinement, the holes are limited in their motion such that they can traverse parallel to the interface but not perpendicular to it, thereby forming a 2DHG. Due to the spatial separation between hole accumulation in the potential well and the acceptor dopant atoms, the effect of ionised impurity scattering is significantly reduced in HHMTs, of benefit to carrier mobility.

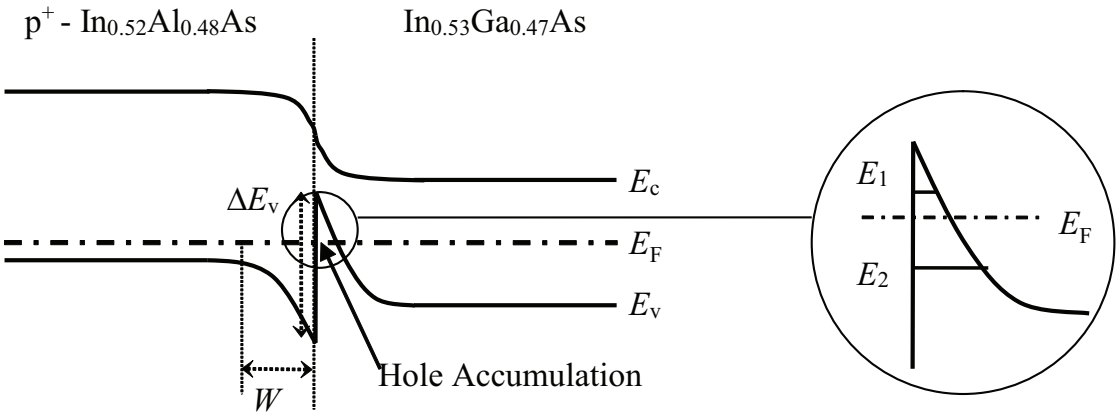


Figure 2.17: Band diagram illustrating heterojunction formation between p-doped $In_{0.52}Al_{0.48}As$ and undoped $In_{0.53}Ga_{0.47}As$ resulting in the creation of a 2DHG. E_1 and E_2 are the subband energy levels.

The above example was based on layers with similar lattice constants. Strain in heterostructures can be obtained by growing a thin material layer of lattice constant a_B on top of another material layer of differing lattice constant a_A. During growth, the in-plane lattice constant of the top layer is adjusted to the lattice constant of the bottom layer, such the two layers possess the same in-plane lattice constant. This is accommodated by a change in the perpendicular lattice constant of the top layer, which is increased for compressive strain and reduced for tensile strain. If $a_B > a_A$, then the top layer is under compressive strain whereas for $a_B < a_A$ the layer is under tensile strain as shown in Figure 2.18. In both cases a biaxial strain is induced in the plane of the layer ($\varepsilon_x = \varepsilon_y = \varepsilon_{||}$) and is given by [108]:

$$\varepsilon_{||} = \frac{a_A - a_B}{a_B}. \tag{2.46}$$

The in-plane biaxial strain also gives rise to uniaxial strain perpendicular to the layer (ε_z=$\varepsilon_{||}$) as follows [108]:

$$\varepsilon_{\perp} = \frac{-2\nu}{1-\nu}\varepsilon_{||}, \tag{2.47}$$

where ν is the Poisson's ratio.

The lattice-mismatch, and thus strain in the layer, can only be accommodated up to a certain thickness known as the critical layer thickness (CLT). Beyond the CLT, the strained layer relaxes with the generation of misfit dislocations, which degrade the channel transport properties. The critical thickness of the strained epitaxial layer can be evaluated using the People-Bean [109] and Matthews-Blakeslee [110] models.

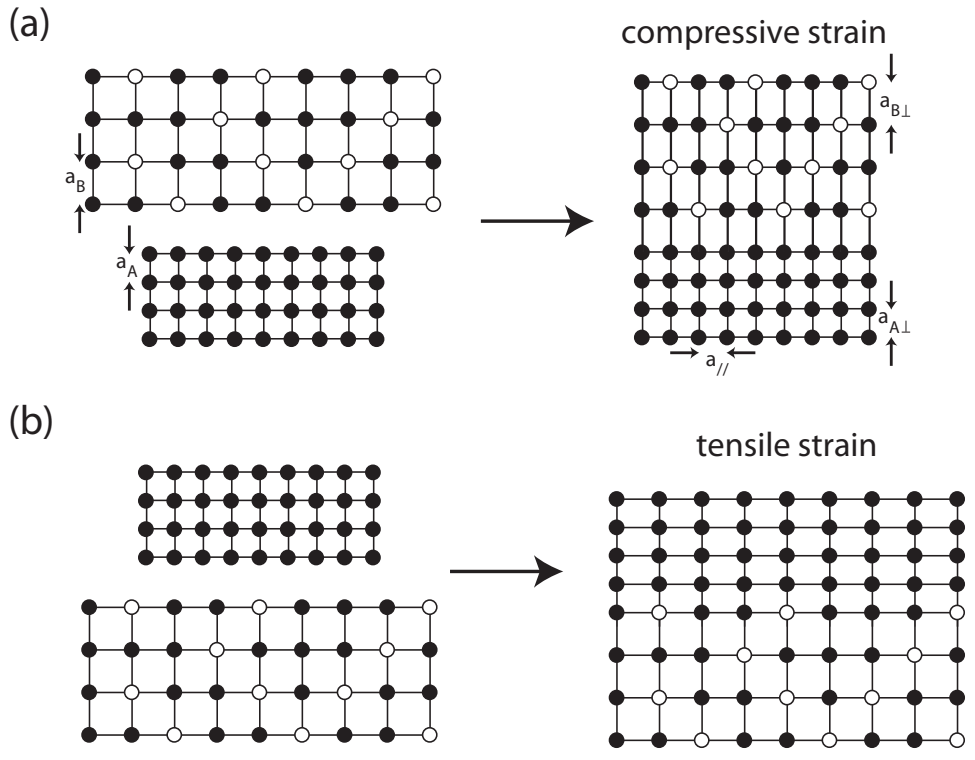


Figure 2.18: Strain due to the growth of a thin material layer of lattice constant a_B on top of another material layer of differing lattice constant a_A: (a) compressive strain for a_B>a_A and (b) tensile strain for a_B<a_A [108].

2.4.2.2 Effects of Strain and Quantum Confinement

A schematic of the bandstructure of unstrained direct bandgap III-Vs is shown in Figure 2.19(a). The valence band comprises heavy-hole (HH) and light-hole (LH) bands which are degenerate in the Brillouin zone centre. This translates to both bands being occupied and having an energy maximum that coincides. The spin-orbit split-off (SO) band is situated far from the valence band maximum, typically between $0.3 - 0.7\,\text{eV}$ for III-Vs [68]. As a result the SO band does not partake in transport. It is noted that the HH and LH band designations refer to the hole masses in the direction perpendicular to the plane.

The effect of biaxial compressive strain on the valence band is twofold and is illustrated schematically in Figure 2.19(b). Firstly, the degeneracy of the HH and LH bands is lifted causing the HH band to move above the LH band. As a result, an increased number of holes will first populate the HH band before beginning to populate the LH band. Secondly, there is a change in the effective mass of holes that populate the bands. Due to the smaller in plane effective mass of the HH band, more holes would possess a lower effective mass. This in turn gives rise to mobility enhancement (see Equation (2.44)).

Strain further induces anisotropy of the bandstructure such that the curvature of the bands can be different in the direction perpendicular and parallel to the growth plane

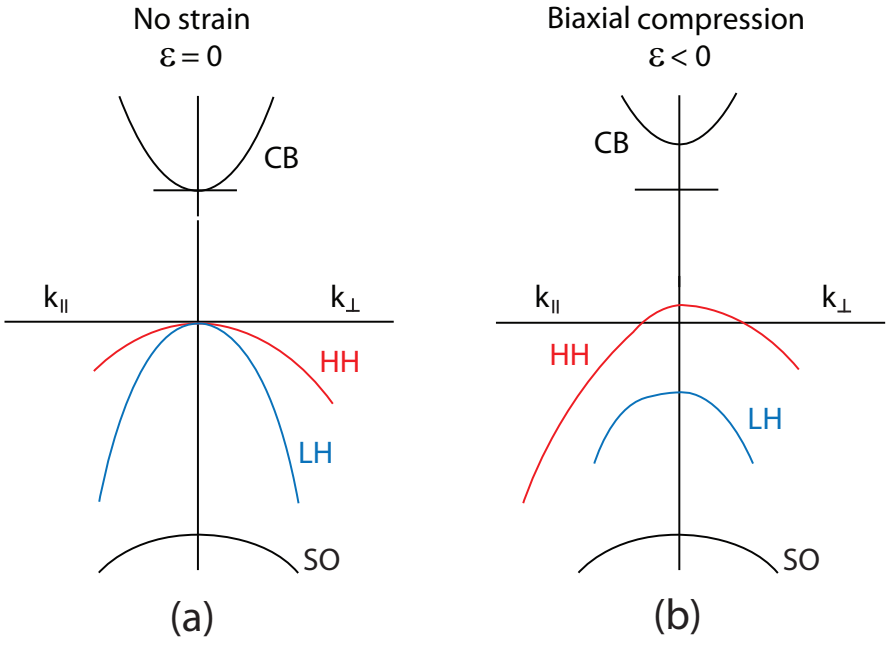


Figure 2.19: Schematic of the bandstructure of direct bandgap III-V semiconductors (a) without strain and (b) with biaxial compressive strain [111].

(Figure 2.19(b)). Therefore the band with the heavier hole mass in the growth direction will possess a lighter mass in the perpendicular direction. This aids in enhanced carrier confinement which is the case for the HH band. The splitting of the bands also causes a change in the valence band energy barrier [111]. The increasing barrier height with strain enhances the charge transfer into the QW, resulting in improved carrier confinement.

Typically strained layers are grown as quantum wells (QWs), whereby a channel is interposed between two barrier layers. The reduction in QW thickness with increasing strain also induces quantum confinement. The effect of quantum confinement is to split the HH and LH bands further into subbands. The effects of both strain and quantum confinement on subband splitting is shown for an $In_{0.83}Ga_{0.17}As/In_{0.52}Al_{0.48}As$ QW for different QW thickness in Figure 2.20 [111]. It is observed that the band splitting between the low in-plane mass HH1 subband and the next subband (HH2 or LH1), of high in-plane mass, increases as the QW thickness is reduced from 20 nm down to 2 nm. As a result, the majority of holes will first populate the HH1 band before occupying the next higher energy. The mobility is therefore increased from more holes having a lower effective mass in the HH1 subband. Furthermore, better carrier confinement is also obtained from the heavier out-of-plane effective mass of the HH1 subband. However there is a limit to which the channel can be thinned with increasing biaxial strain since the effects of interface scattering become more pronounced, resulting in mobility degradation.

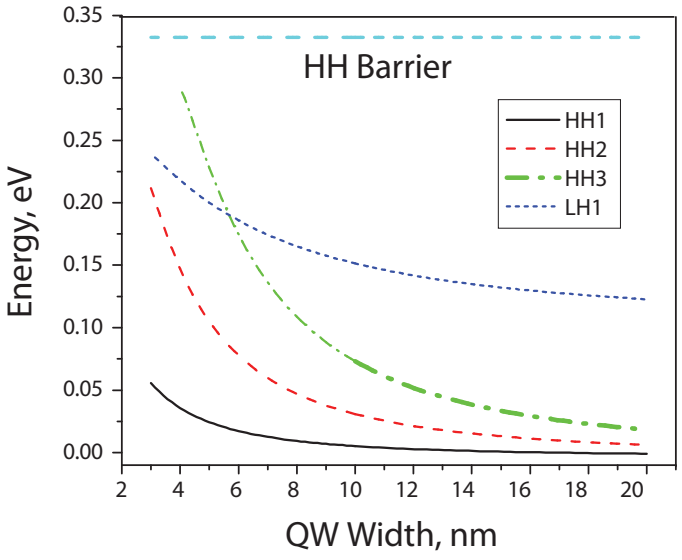


Figure 2.20: Plot of energy at $k = 0$ vs. QW thickness for subbands in a $In_{0.83}Ga_{0.17}As/In_{0.52}Al_{0.48}As$ QW. Higher energy corresponds to a smaller barrier [111].

2.5 Metal-Semiconductor Contacts

Parasitic components are a major performance limiting factor in field-effect-transistors (FETs). From a DC perspective, parasitic series resistance associated with S/D regions, R_{sd}, outwith the intrinsic region (gated) can have a sizeable contribution to the overall resistance given by:

$$R_{tot} = R_{ch}(V_g) + R_{sd}, \tag{2.48}$$

where $\mathrm{R_{ch}(V_g)}$ is the gate-controlled channel resistance and R_{sd} is the sum of the source resistance (R_{s}) and drain resistance (R_{d}), as shown in Figure 2.21.

The drain-source and gate-source voltages dropped across the intrinsic region (intrinsic biases) - $\mathrm{V_D}$ and $\mathrm{V_G}$ - and the applied drain-source and gate-source voltages (extrinsic biases) - $\mathrm{V_d}$ and $\mathrm{V_g}$ - are connected through the following equations [112]:

$$V_d = V_D + I_d(R_s + R_d), \tag{2.49}$$

$$V_g = V_G + I_d R_s, \tag{2.50}$$

where $\mathrm{I_d}$ is the current flowing through the transistor. It is evident from Equations (2.49) and (2.50) that the parasitic resistances cause the intrinsic biases to incur a drop in voltage from the extrinsic biases. The effect of these voltage losses on the MOSFET output characteristics is two-fold; a reduction in $I_{\mathrm{d,sat}}$, and a drop in the linear region output conductance owing to the decrease in the gradient of I_{d}-V_{d}. Mitigating the effect of parasitics becomes increasingly important as gate lengths are scaled to smaller dimensions. The increase in on-current (I_{on}) expected from a decrease of the channel resistance with scaling can only be realised if R_{on} is not dominated by R_{sd} [113].

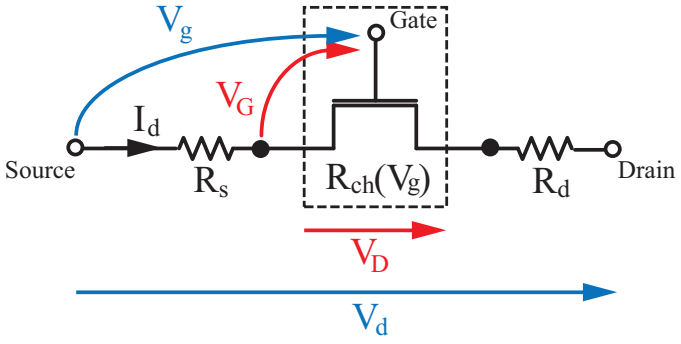


Figure 2.21: Schematic illustration of the source/drain series resistances (R_{s}, R_{d}) of a FET and the biases, both extrinsic (V_{g}, V_{d}) and intrinsic (V_{G}, V_{D}), associated with the device.

Another figure-of-merit that is impacted by parasitic series resistance is the measured (extrinsic) transconductance, g_{m}. Following from Equation (2.50), $g_{\mathrm{m}}\left(=\frac{\delta I_d}{\delta V_g}\right)$ is related to the intrinsic $\mathrm{g_{mi}}\left(=\frac{\delta I_d}{\delta V_G}\right)$ through the following relationship [112]:

$$g_m = \frac{g_{mi}}{1 + g_{mi}R_s}, \tag{2.51}$$

which takes account of the fact that a fraction of any increase in the extrinsic gate bias appears as an increased voltage drop across the source resistance, R_{s}. The drive current modulation efficiency suffers from the effect of R_{s}, which acts to lower the measured g_{m} from its intrinsic value. For large values of the source resistance whereby $\mathrm{g_{mi}}R_{\mathrm{s}} >> 1$, the extrinsic transconductance becomes independent of gate bias and is reduced to $1/R_{\mathrm{s}}$.

In this section, the limiting cases for the metal-semiconductor interface are first introduced, before mechanisms of current transport across the interface are analysed.

2.5.1 Metal-Semiconductor Interfaces

The metal-semiconductor (M/S) interface has two main limiting cases; the Schottky-Mott limit [114] and the Bardeen limit [115]. These limiting cases are considered in the following paragraphs.

When a metal and a semiconductor, whereby $\phi_{\mathrm{m}} < \phi_{\mathrm{s}}$, are brought into contact, the Fermi levels align to maintain equilibrium. This results in the diffusion of holes from the semiconductor to the metal. The build up of positive charge on the metal balanced by negatively charged acceptor ions exposed in the semiconductor generates an electric field at the M/S interface and induces band bending in the semiconductor. This in turn creates an energy barrier or Schottky barrier to hole transport from the semiconductor to

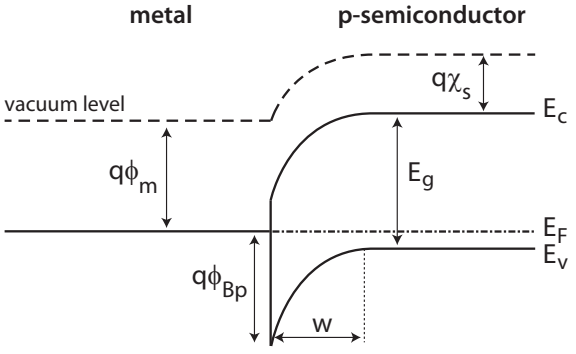


Figure 2.22: Energy band diagram of ideal metal/p-doped semiconductor Schottky contact.

the metal as shown in Figure 2.22. In the Schottky-Mott limit, the height of the barrier (ϕ_{Bp}) is given by Equation (2.52):

$$q\phi_{\mathrm{Bp}} = E_{\mathrm{g}} - q(\phi_{\mathrm{m}} - \chi_{\mathrm{s}})\,. \tag{2.52}$$

In this case, the barrier height can be tuned by the choice of gate metal and the doping in the semiconductor. This model assumes an ideal contact, free of surface states.

In practice, many semiconductor system contain surface states and therefore obey the Bardeen limit, which accounts for surface states. This is also the more relevant model for understanding contact formation to III-V systems. The high density of interface traps in III-V systems pins the Fermi level above the valence band. This in turn renders the energy barrier independent of the metal work function, an example of which is illustrated in Figure 2.23 for the case of GaAs. For contacts exhibiting a pinned Fermi level, the barrier height is determined by the surface state density and the semiconductor bandgap.

The width of the depletion region formed by the barrier can be calculated using Poisson equation and is given as:

$$w = \sqrt{\frac{2\epsilon_{\mathrm{s}}\epsilon}{qN_{\mathrm{A}}}(\phi_{\mathrm{Bp}} - V)}\,, \tag{2.53}$$

where w is the barrier width and V is the applied voltage.

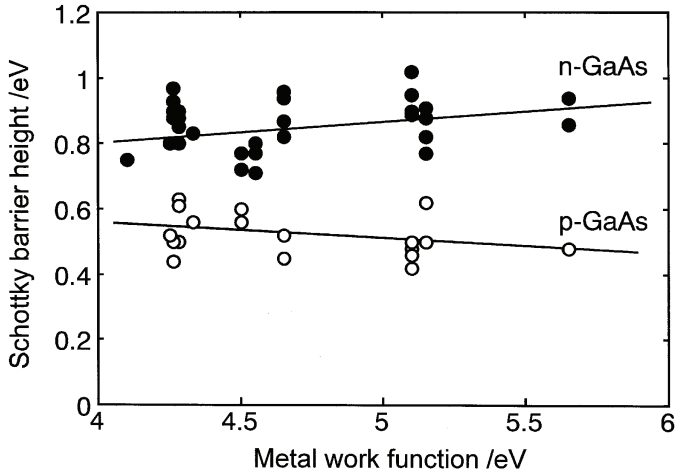


Figure 2.23: Measured Schottky barrier heights (ϕ_{B}) for metals of a variety of work functions on n-type and p-type GaAs [116].

2.5.2 Current Transport Mechanisms

In order for current to flow, holes need to overcome the energy barrier imposed by the M/S interface. Current transport across the M/S interface is governed by three principal mechanisms; thermionic emission (TE), field emission (FE) and thermionic-field emission (TFE). The dominant mode of transport is determined by the barrier width and height, which as shown in Equation (2.53), is dependent on the doping density and the intrinsic material properties. These transport mechanisms are described in the following sections.

Thermionic emission (TE)

Thermal excitation of holes across the energy barrier occurs in relatively low doped semiconductors (N_A<10^{17} cm^{-3}). At such doping levels, the depletion region is too wide for holes to directly tunnel through (Equation (2.53)). As a result, holes must acquire sufficient thermal energy to overcome the barrier and the resulting TE current is expressed as:

$$J_{TE} \propto e^{E_B/kT}, \tag{2.54}$$

where the energy barrier height, E_B, takes on different values depending on the bias condition as illustrated in Figure 2.24. In thermal equilibrium, the barrier height is just the built-in potential (qV_{bi}). When a reverse bias (V_r) is applied, the Fermi level is brought closer to the energy needed for holes to transit the barrier. As a result, the barrier height is decreased to $q(V_{bi}-V_r)$, which in turn enhances thermal excitation of holes. Under a forward bias (V_f), however, the barrier height impeding current flow from metal to semiconductor is just $q\phi_{Bp}$, as dictated by the surface states in the case of III-Vs. For a small barrier height, the likelihood of holes surmounting the barrier by thermal excitation increases and the resulting contact is ohmic. Alternatively, if the barrier height is large,

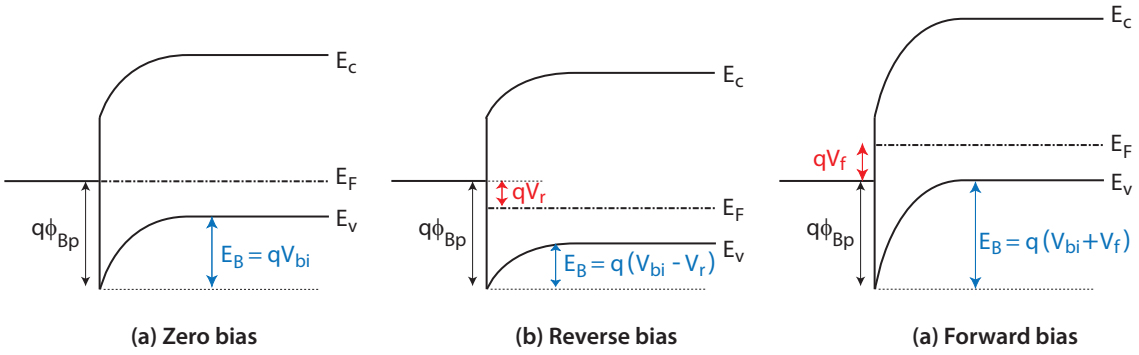


Figure 2.24: Energy band diagrams of metal/low p-doped semiconductor Schottky contact under various bias conditions.

only a small fraction of holes will overcome the barrier, resulting in a minimal forward bias current at low bias. This gives rise to a rectifying (Schottky) contact. With increasing V_{f}, the depletion regions decreases and at a certain bias breakdown will occur by the process of FE as described in the next section. As seen from Equation (2.54), thermionic emission is enhanced by increasing the temperature and by employing semiconductors with a small bandgap, thus reducing the barrier height.

Field Emission (FE)

As doping in the semiconductor is increased, the width of the barrier reduces (see Equation (2.53)). Consequently, quantum mechanical tunnelling of holes directly through the barrier becomes possible. For heavily doped semiconductors (N_{A}>$10^{18}\,\mathrm{cm}^{-3}$), the width is reduced to such an extent that appreciable tunnelling can occur and thermal excitation of holes over the barrier may be negligible. Field emission has little dependence on temperature since only holes near the Fermi level are involved in the tunnelling process. Figure 2.25 illustrates the energy band diagram for a field emission process. The resulting tunnelling current is described in Equation (2.55):

$$J_{\mathrm{FE}} \propto e^{-q\phi_{\mathrm{Bp}}/E_{00}}\,, \tag{2.55}$$

where E_{00} represents the hole tunnelling probability:

$$E_{00} \propto \frac{qh}{4\pi}\sqrt{\frac{N_{\mathrm{A}}}{m^{*}\epsilon_{\mathrm{s}}\epsilon}}\,. \tag{2.56}$$

As evident from Equations (2.55) and (2.56), the tunnelling current shows a strong dependence to doping density, and thus barrier width. Therefore, the probability of tunnelling increases as the doping increases, which is concomitant with thinning of the barrier.

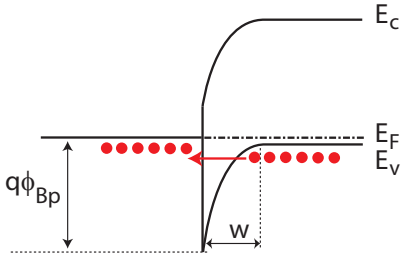


Figure 2.25: Energy band diagram of metal/heavily p-doped semiconductor Schottky contact.

Thermionic-Field Emission (TFE)

There is also a doping range ($10^{17}\,\mathrm{cm}^{-3}<N_\mathrm{A}<10^{18}\,\mathrm{cm}^{-3}$) for which thermionic and field emission combined provide for current flow at the M/S interface. At these densities, due to too wide a barrier, direct tunnelling is not possible and holes do not possess sufficient thermal energy to cross the barrier. However, holes at elevated energies can still tunnel through the barrier, as its width narrows with increasing energy (Figure 2.26). The resulting current due to TFE is expressed as:

$$J_\mathrm{FE} \propto \exp\left(\frac{q\phi_\mathrm{Bp}}{E_{00}\coth(\frac{E_{00}}{kT})}\right), \tag{2.57}$$

where E_{00} is given by Equation (2.56). In this case, the current is dependent on both the doping density and barrier height, and is expected to increase with temperature.

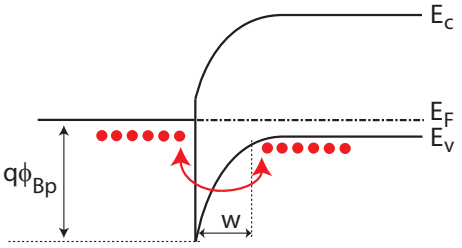


Figure 2.26: Energy band diagram of metal/moderately p-doped semiconductor Schottky contact.

3

Fabrication Techniques

3.1 Introduction

Device fabrication is an involved process that draws on a variety of techniques. Compatibility of these techniques, with the substrate and pre-existing features, should be taken into account when devising process flows in order to realise devices of high performance, yield and reliability. A fundamental understanding of the principles of fabrication and the associated processing techniques is therefore vital. This chapter provides an overview of the fabrication techniques that were used as part of this work; lithography, epitaxial material growth, dielectric and metal deposition, and etching. These techniques are presented and discussed both from a general perspective and as applicable to III-V MOS device processing, taking into consideration the capabilities and limitations associated with each method.

3.2 Lithography

In microelectronics, lithography is the process by which a pattern is transferred onto a substrate. A range of lithographic techniques such as optical lithography, electron beam lithography (EBL), ion-beam lithography and imprint lithography can be employed for this purpose. These techniques are differentiated by their pattern transfer action, e.g. the application of photons in optical lithography in contrast to pressure force in imprint lithography. Although optical lithography was used at times, the bulk of the fabrication work of this project was underpinned by EBL.

3.2.1 Lithography Basics

Most lithographic techniques used in pattern transfer are based on the same fundamental concepts. The process of pattern transfer begins with uniformly spin coating the substrate with a polymeric radiation-sensitive film, referred to as resist. Subsequently, the resist is selectively irradiated. This leads to an alteration of the physical or chemical properties of the resist in the exposed region in terms of its solubility, dictated by the type of resist used. In the development stage that follows, a suitable solvent is used to remove the resist in a selective manner.

Two types of resist tones, positive and negative, can be employed for pattern transfer. The terms positive and negative are a reflection of the change in resist solubility following irradiation. In positive resists irradiation causes chain scission [117], a polymer degradation mechanism which gives rise to enhanced resist solubility. Dissolution of the resist in the exposed region during the development step leaves the substrate masked in the unexposed region. On the other hand the cross-linking polymer reaction [117], that ensues in irradiated negative resists, diminishes the resist solubility. As a result the substrate is only masked in the exposed region, due to resist dissolution in the unexposed region, during development. Figure 3.1 schematically illustrates the different final resist profiles achieved when from using positive and negative resists. The selectively masked resist template is subsequently exploited to define the final features during pattern transfer. This process is described in Section 3.3.

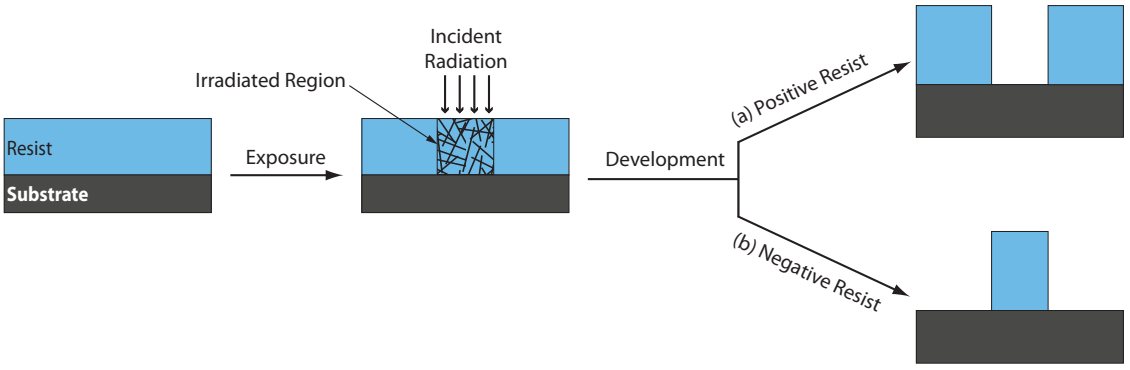


Figure 3.1: Resultant resist profiles for positive and negative resists during pattern transfer.

3.2.2 Optical Lithography

Optical lithography has remained a patterning mainstay in microelectronics over the last three decades. In its simplest form a light source, typically ultraviolet (UV), is irradiated

onto a photo-sensitive resist through a physical mask that is in direct contact with the substrate or at a small distance away from it. The former is known as contact lithography (Figure 3.2(a)) while the latter is referred to as proximity lithography (Figure 3.2(b)). The mask plate, usually made of quartz, comprises transparent and opaque (chromium-coated) regions that define the patterns to be written.

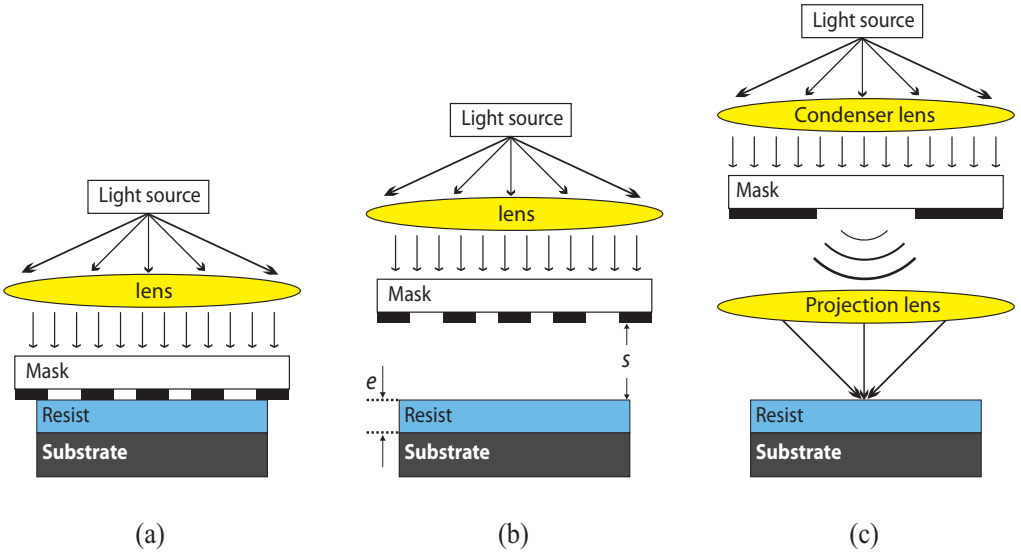


Figure 3.2: Schematic illustration of optical lithographic techniques: (a) contact lithography, (b) proximity lithography and (c) projection lithography.

In optical lithography, light gets diffracted at the edges of the opaque regions of the mask thereby limiting the achievable resolution. The resolution limit of contact and proximity lithography is given by:

$$l_{\min} = \frac{3}{2}\sqrt{\lambda(s + \frac{1}{2}e)}\,, \tag{3.1}$$

where λ is the radiation wavelength, s is the mask-to-substrate separation, and e is the resist thickness [118]. Although contact lithography is capable of high resolution, both mask and substrate are subject to damage from abrasion and debris during the contacting process, which in turn compromises pattern fidelity. In addition, contacting surfaces should be perfectly planar for ideal contact. This however is not always possible as the resist may not be able to smooth out the topographical relief on the substrate from previous processing. The potential debris on the mask further adds to the problem of non-ideal contact [118]. This affects lithographic alignment and introduces a finite, non-uniform mask-to-substrate separation which impacts resolution. While resolutions as small as $0.2\,\mu$m are achievable using thin resists exposed at short wavelengths, e.g. deep ultraviolet (DUV)

light, this is often limited to about 1 μm in laboratory-based lithography tools operating at longer wavelengths (typically 400 nm) and affected by the aforementioned non-idealities. In contrast, mask and substrate damage is circumvented in proximity lithography albeit at the expense of resolution. An offset of 10 μm is usually sufficient to ensure consistent separation is maintained along all points between the mask and the substrate, which then limits the minimum resolution to about 3 μm. The simplicity of implementation and low cost associated with these techniques have made them a permanent feature in research and university laboratories.

In industry, projection lithography has been adopted as the aforementioned techniques are not suited to meet the demands of feature size scaling or mass production. This technique relies on a projection lens that focuses and reduces the mask patterns prior to imaging onto the resist, with the mask held at some distance away from the substrate (Figure 3.2(c)). Its resolution limit is governed by the Rayleigh equation:

$$l_{\min} = \frac{\kappa . \lambda}{\mathrm{NA}}, \tag{3.2}$$

where the scaling factor κ is a measure of the difficulty associated with the lithographic process in printing a given feature and NA is the numerical aperture of the projection optics, which is a measure of the light gathering ability of the lens [118–120]. Historically the main enablers of resolution enhancements, in line with Equation (3.2), have been the transition to shorter wavelengths and reduction in the scaling factor through improvements in the illumination, quality of optics and photoresist performance. These advances were further complemented by various resolution enhancement techniques such as phase-shift masking, proximity optical correction, off-axis illumination and more recently double patterning [118, 120]. Notably in 2003 the 100 nm feature size barrier was broken with the mass production of 90 nm gate length transistors [119].

In addition to the capacity to down-scale feature sizes, projection lithography's [119, 120] other attributes of high throughput, extremely low defects and high overlay capability, all vital requirements of high-volume chip manufacturing, have collectively cemented its dominance in microelectronics. The scaling trend towards nanometric features continues to progress with the aid of liquid-immersion lithography. Resolution enhancement is achieved by increasing the refractive index of the medium between the projection lens and the resist by using a fluid in place of air, which in turn increases NA [119, 120]. This technology is expected to be extended down to the 11 nm node before next generation techniques such as extreme ultraviolet (EUV), imprint or maskless lithography are introduced.

3.2.3 Electron Beam Lithography

While feature size scaling into the sub-100 nm regime is a fairly recent development in optical lithography, features as small as 20 nm have been demonstrated using EBL as far back as the 1980s [121]. EBL uses a finely focussed beam of electrons to scan and expose the resist-coated substrate during pattern writing. This direct writing approach does away with physical masks as the process of beam deflection and blanking, for pattern exposure, is controlled by a computer system that reads in software-generated pattern layouts. As a result, EBL is a versatile technique that affords the latitude of changing the resulting patterns as often as necessary. This is especially advantageous when the nature of work is highly exploratory as is often the case when developing new processes. Moreover, degradation of patterns arising from mask defects is alleviated.

In optical lithography diffraction of light limits the achievable resolution, as highlighted in the preceding section. However, an electron beam (e-beam) is generally not diffraction-limited. This is because the de Broglie wavelength of the e-beam is of the order 4–12 pm at accelerating voltages of 10–100 kV during operation. Thus, the main advantage of EBL is the high resolution patterning capability it offers. On the other hand, the serial writing of patterns makes for a slow process, therefore from a throughput and cost perspective it is not beneficial to use EBL for industrial scale production of highly dense, well established patterns. In research and development however, the other features of EBL processing outweigh throughput.

There are a number of factors that determine the resolution, sensitivity, accuracy and reliability of pattern transfer in EBL. The resolution is influenced by the spot size of the e-beam, which is determined by the quality of the electron-optics. This is further complicated by scattering resulting from electron-resist and electron-substrate interactions during exposure. In addition the substrate material, resist and developer used, together with the EBL processing parameters, all have a profound effect on pattern definition. Finally, the accuracy of pattern overlays is determined by the alignment techniques adopted. These different aspects of EBL are discussed in the following sections.

3.2.3.1 E-Beam Apparatus and Operation

At the University of Glasgow, a Vistec VB6 Ultra High Resolution Extra Wide Field (UHR EWF) EBL tool is used for pattern exposure. Although the description provided herein is based on the VB6 tool, the main aspects presented are common to most EBL systems. The VB6 tool comprises of a number of principal components, as illustrated in

Figure 3.3. At the heart of the system is the electron-optical column (EOC), which is used to form and direct an e-beam onto a resist-coated substrate, with control over the beam intensity, focus and deflection. Detector electronics, that provide useful information about the beam, and a high precision stage, on which the substrate is mounted, are located at the end of the column. In addition to providing power and temperature control, ultra-high vacuum of the tool, necessary for high-resolution patterning, is maintained by the support system. Critical to EBL operation is the pattern generator that converts software generated pattern layouts into analogue control signals for beam deflection and blanking.

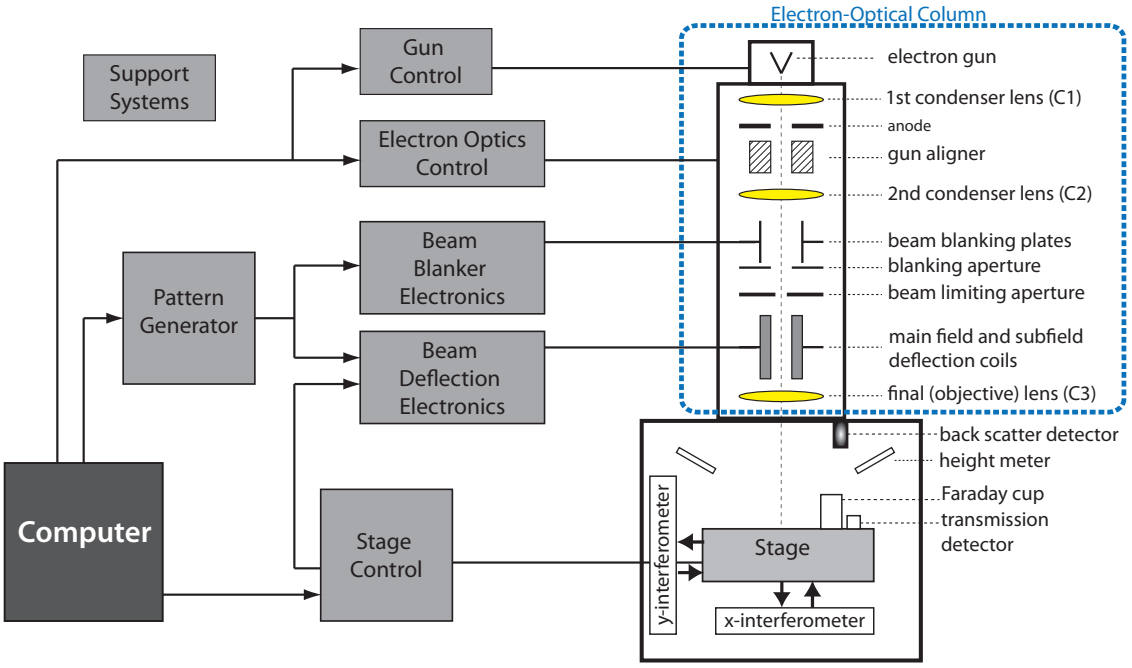


Figure 3.3: Schematic illustration showing the major components of the Vistec VB6 UHR EWF electron beam lithography system.

Electron-Optical Column (EOC)

The EOC is configured such that the electron gun (source), from which the e-beam originates, sits at the top-tier of the column. All other elements responsible for beam modification to produce a spot of desirable size and current on the substrate are contained within the bottom-tier. These include electron optics, apertures, beam blanking and deflection units.

Of significance to high-resolution patterning is the ability to obtain spots with fine focus. Although the beam may start off small from the source, aberrations associated with imaging and beam deflection tend to enlarge the beam as it travels down the column. The various aberrations affecting the beam and its corresponding definitions are given

in Table 3.1. A large angle of beam convergence, originating from the lenses and deflectors, gives rise to spherical aberration. The resulting spot is de-focussed and distorted. However, constricting the beam convergence can result in diffraction effects becoming more noticeable, albeit its contribution is significantly small as previously highlighted. Broadening of the electron energy distribution gives rise to chromatic aberration. This can arise in the source but can have contributions from the space-charge effect, resulting from electron-electron repulsions at beam crossovers. High beam currents compound this effect, which further acts to blur and de-focus the spot. Therefore, the final spot that is imaged onto the substrate, taking into account the aberration contributions, is expressed as [122]:

$$d_{\text{min}} = \sqrt{d_{\text{g}}^2 + d_{\text{s}}^2 + d_{\text{c}}^2 + d_{\text{d}}^2}\,. \tag{3.3}$$

The following description of the different elements of EOC takes into consideration these aberrations where d_{g}, d_{s}, d_{c} and d_{d} are defined in Table 3.1 below.

Table 3.1: List and definition of parameters contributing to final spot size, d_{min}. Adapted from [122].

• beam spot without aberration, $d_{\text{g}} = \frac{d_{\text{v}}}{M}$	d_{v}, virtual source size; M, demagnification
• spherical aberration, $d_{\text{s}} = \frac{1}{2} C_{\text{sa}} \alpha^3$	C_{sa}, spherical aberration coefficient; α, beam convergence angle
• chromatic aberration, $d_{\text{c}} = C_{\text{ca}} \alpha \frac{\Delta V}{V}$	C_{ca}, chromatic aberration coefficient; ΔV, beam energy spread; V, beam acceleration voltage
• diffraction limit, $d_{\text{d}} = 0.6 \frac{\lambda}{\alpha}$	$\lambda = 1.2/\sqrt{V}$, electron wavelength

In the source, the beam is formed via a cathode emission process. Due to its good emission characteristics a thermionic field emitter (tungsten needle coated with zirconium oxide) is used as the cathode filament in the VB6. The small "virtual" source size (d_{v}) of 15–20 nm [123] allows for a smaller beam spot (d_{g}) to be formed on the substrate with less demagnification and fewer lenses. In addition, the low electron energy dispersion (ΔV) of 0.9 eV [124] reduces chromatic aberrations (d_{c}), this in turn minimises beam broadening. These source characteristics enable higher resolutions to be achieved. On the other hand, the high source brightness provides for a high beam current, which then requires less dwell time. This is because the product of the two parameters sets the dose for a given exposure as:

$$D = \frac{I_{\text{b}}\, t_{\text{b}}}{S_{\text{b}}^2}\,, \tag{3.4}$$

where D is the exposure dose in $\mu\text{C/cm}^2$, I_{b} is the beam current in nA, t_{b} is the length of time the e-beam is still at each location during exposure (dwell time) in μs and S_{b}^2 is the area exposed at each point where the beam step size (S_{b}) is the distance the beam is

deflected between exposures in nm [124]. The increased throughput comes at the expense of resolution, due to the space-charge effects mentioned earlier. A number of electrodes contained within the gun extract and accelerate the electrons from the filament, which is further accelerated by the anode to the specified beam acceleration voltage, which typically is 50 kV or 100 kV in the VB6. Higher acceleration voltages further reduce chromatic aberrations and space-charge effects. The emergent beam is aligned to the optical axis of the lenses by the gun alignment system comprising of magnetic deflection coils which provide tilt and shift correction.

One electrostatic (C1) and two magnetic (C2, C3) lenses are contained within the column. C1 is located at the source and is used for initial beam focus, following which the beam is further focussed by C2. The focus point on C3 (objective lens) is fixed by C2. This configuration caters for easy adjustment of the spot size on the substrate without having to change the beam focus or current density. Thus, stability and repeatability are achieved during pattern writing. The crucial beam focus on the substrate is provided by C3, and a tight focus with minimum aberrations is achieved by minimising the distance between this lens and the substrate (35 mm in the VB6). For critical focussing, magnetic lenses are preferred to electrostatic lenses since the former create fewer aberrations.

Turning the beam on and off, referred to as blanking, is achieved with a pair of plates configured as an electrostatic deflector this provides for high-speed beam interruptions. The degree of beam deflection is controlled by the voltage applied to the plates. A large enough angular deflection causes the beam to be interrupted by an aperture, thereby turning the beam off. The column is arranged such that the electrical centre of the blanker coincides with the beam crossover created by lens C2. This configuration, known as conjugate beam blanking, ensures that the beam position on the substrate is independent of the blanker voltage up to the point of the beam turning off. As a result any unintended exposure arising from beam movements during the blanking process is avoided.

As the beam traverses the column, it encounters a series of apertures that have designated purposes. Spray apertures serve to shield the substrate from stray electrons generated in the column and exclude beam electrons that are a long way off the optical axis. Based on the exposure requirements, beam limiting apertures set the size and current of the beam on the substrate. Smaller apertures provide a larger constriction of the convergence angle, thereby minimising lens aberrations. The increased resolution comes at the expense of usable beam current. As a result, a larger dwelling time is required for a given exposure, which reduces throughput. Furthermore, blanking apertures provide for beam interruptions when used in conjunction with the beam blanker as previously

highlighted.

The process of pattern exposure involves forming the beam, setting its size and current, and controlling its relative motion to the intended position on the substrate. The latter is performed by the main field and subfield deflection coils. Although electrostatic deflections have a faster response they introduce more aberrations and thus these coils are magnetically driven. In addition, large angular deflections from the optical axis give rise to increased aberrations affecting the beam. This is circumvented by limiting the maximum distance within which the beam can be deflected. This is referred to as the main field, for which deflection is provided by the main field deflection coils. The large angles of deflection of such coils limit the speed of writing, therefore subfield deflection coils, that have smaller angular deflections and therefore a faster writing time, are used within the main field. This two-tier deflection system is implemented by dividing up the main field into a 64×64 array of smaller fields, known as subfields. This is a process done by the pattern generator hardware in real-time. Exposure is then achieved by first moving the beam via main field deflection to the centre of a subfield, following which subfield deflection is used for scanning the beam over the individual pixels comprising the subfield. This sequence is repeated for all subfields within a main field.

The main field size is determined by the maximum deflection range as limited by the precision of the digital-to-analog converter (DAC) and the resolution of the pattern generator. A 20-bit pattern generator is installed in the VB6 with 1.25 nm, 1.0 nm and 0.5 nm resolution capabilities. As such, at the maximum resolution setting, the main field size comes out to 1.31 mm with subfields in the order of 20.5 μm. The size of the main field and its corresponding subfields decrease with increasing resolution setting.

A number of other components, located within the EOC, provide crucial information about the beam and are used for various purposes during EBL operation. A Faraday cup and transmission detector are mounted on the stage to measure the beam current and spot size respectively. Height variations between the objective lens and the substrate can introduce errors in the form of field size and focus variations. To overcome such errors a height meter, comprising of a charge coupled device (CCD) array detector and an infrared laser, is included in the column. Height readings are obtained by detecting the position of the laser beam reflected off the substrate surface after every stage movement. This information is then used to automatically compensate for errors, which in turn maintains the accuracy of beam deflections. The final component is the back scatter electron (BSE) detector mounted on the bottom of the objective lens, this provides a means of imaging the substrate as the e-beam is scanned across the substrate. This is useful for calibrating

beam deflections, focus control and detecting alignment markers, this is described further in Section 3.2.3.4.

Writing Strategy

The nature of pattern writing adopted in an EBL tool is determined by the beam shape, beam scanning method and stage control, concomitantly.

The most important distinction between EBL tools arises from the shape of beam employed. The description of the beam thus far has been based on a Gaussian beam, which is the smallest spot size that is focussed onto the substrate with a Gaussian intensity distribution. Exposure proceeds in a serial manner with each pixel exposed in turn, which results in reduced throughput. The alternative is the shaped beam, which provides an area exposure of broad, uniform beam intensity after passing through a series of beam-shaping apertures. The extension of the beam from a spot to an area results in the exposure of several pixels in parallel. This results in the advantage of increased throughput but also results in reducing the achievable resolution.

The beam can be scanned across patterns in two ways; raster scan or vector scan. In the former, the beam is continuously deflected across a subfield as line scans. Along the line scan, the beam is blanked in un-patterned areas, while patterned areas are exposed. After each line scan the writing is suspended to allow the deflectors to be repositioned to the start of the next line to be scanned. Therefore, the writing time is dependent only on field size and not on pattern density, however the dose variations that can be applied are limited with this method. In the vector scan method the beam is deflected to the patterned areas for exposure, and blanked as it is deflected from one pattern to the next, skipping over un-patterned areas. The writing time therefore is directly proportional to density of patterns, but the advantage is the ability to vary the dose between patterns. Thus, for sparsely packed patterns a vector scan method is more beneficial.

As previously mentioned, the maximum allowed deflection is limited to the main field size. Pattern layouts however are typically larger than this field, in which case stage movements are needed. In the step-and-repeat approach, after the complete writing of a field, the writing is stopped and the stage is moved to the next field for which the process repeats. This sequence of field-stepping enables patterns to be written while the stage is stationary. However, the associated settling periods needed after each stage movement reduces throughput. In contrast, for the moving stage approach, a scan similar to a raster scan ensues as the stage is moved continuously in one direction. At the end of the scan, writing is stopped to enable the stage to be moved to the start of the next raster to be

scanned. This method provides for a higher throughput and like the raster scan method, writing time is independent of pattern density.

Stitching is the process of combining main fields together to realise the overall pattern layout, for which inaccuracies can arise at the field boundaries during translational movements. Due to mechanical limitations of the stage, discrepancies of a few μm between the measured and desired stage positions can occur after each stage movement. Such stitching errors are rectified by repositioning the beam to the desired position on the substrate through beam error feedback (BEF) correction. In this method, the absolute stage position is measured interferometrically in two axes, to a resolution of 0.6 nm, which is then fed back to the beam deflection coils for realigning the beam accordingly. In view of stitching errors, patterns of critical geometries are not placed at field boundaries.

The VB6 is a vector scan system in which patterns are written using a Gaussian beam, with a minimum spot of 3-4 nm, in a step-and-repeat exposure.

Pattern Preparation

Generating the pattern file in software, to be used by the pattern generator's hardware to provide the requisite control signals to the EOC for exposure, is a three stage process which involves pattern design, fracture and layout.

Computer aided design (CAD) packages such as Tanner L-Edit are used in the design stage to create and arrange patterns into different layers. The resultant output files, in GDSII format, are then used in the pattern fracture stage. The process of fracturing is needed to convert the designed data files into basic shapes the pattern generator is capable of handling for e-beam exposure which in this case are trapezoids. In addition to this, data is arranged into subfields and fields as defined by the pattern resolution and maximum deflection limit set by the pattern generator. Fracturing is achieved using the CATS software from Synopsys. The final step that is needed before exposure is the arrangement of the fractured data files into a pattern layout as required, as well as setting of exposure parameters. The software, beam-writer exposure layout for lithographic engineers (BELLE), created by Dr. Stephen Thoms (University of Glasgow) is used for this purpose.

3.2.3.2 Electron-Matter Interaction and the Proximity Effect

A small beam spot size is important for high resolution EBL, however it is not the main resolution limiting factor since modern EBL tools can achieve minimum spot sizes of a few

nm. The effects of scattering phenomena in the resist and the substrate during irradiation ultimately limit the achievable feature size.

Two types of scattering processes, elastic and inelastic, occur when primary electrons (incident e-beam) interact with a resist-coated substrate. Elastic scattering results from electron-nucleus interaction whereas inelastic scattering is due to electron-electron interaction. The latter causes excitation and/or ionisation of atoms, and in the process generates secondary electrons. The majority of these secondaries tend to be typically low in energy (few tens of eV). As a result, their lateral straggle is limited to several nm in range. However, secondary electrons of much higher energies (a few keV) can be significant in number. These fast secondaries have a longer lateral range of tens of nm. The resultant energy distribution (exposure) within the resist is determined by the energy transferred from inelastic scattering, for which secondary electron contribution is significant [124].

During e-beam irradiation the majority of the primary electrons that penetrate the resist undergo inelastic scattering, which leads to resist exposure. The small-angle deflection of primary electrons that follows from inelastic scattering, known as forward scattering, broadens the e-beam laterally in the resist. With a large number of electrons penetrating into the substrate, forward scattering broadens the beam further in the substrate. Some of these electrons are also elastically scattered due to collisions with heavy substrate nuclei, the result of which is large-angle electron scattering. This is known as back scattering. Back scattered electrons arrive back into the resist at distances far away from the original point of e-beam incidence, leading to undesired resist exposure [123]. These scattering processes are outlined in Figure 3.4(a). The scattering-specific energy distributions in the

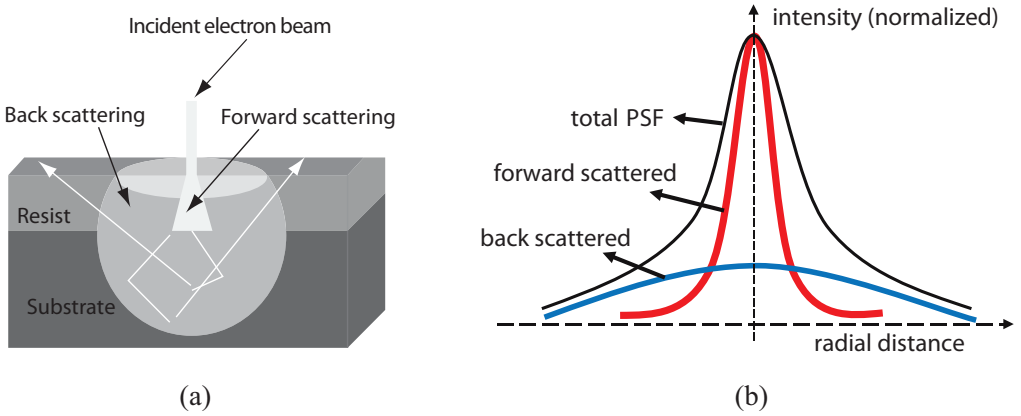


Figure 3.4: Schematic depicting (a) the scattering processes arising from electron interaction with the resist and substrate, and (b) Gaussian energy distribution profiles of the individual scattering processes together with the total point spread function (PSF) of their combined distributions.

resist are Gaussian approximated. Forward scattering distribution tends to be narrower while back scattering distribution is broader, and their summation gives form to the overall energy distribution profile, known as point spread function (PSF). These distribution profiles are illustrated in Figure 3.4(b).

The relative scattering contributions to the PSF are dependent on a number of process conditions as illustrated in Figure 3.5. Thinner resists and higher incident e-beam energies reduce the deflection angle of primary electrons, therefore the extent of beam broadening due to forward scattering is reduced (Figures 3.5(a) and (b)). The back scattered distribution is also affected by the beam energy. Increasing beam energies deepen and widen the back scattering effect, the result of which is increased distribution width along with reduced intensity (Figure 3.5(b)). This leads to the back scatter range extending to a larger radial distance, for which effective exposures are broadly uniform and diluted. In addition, substrate density influences the back scattering effect. Higher atomic number substrates, such as GaAs, result in an increased number of scattering events. This in turn increases the intensity but the back scatter range is reduced (Figure 3.5(c)) [122, 124].

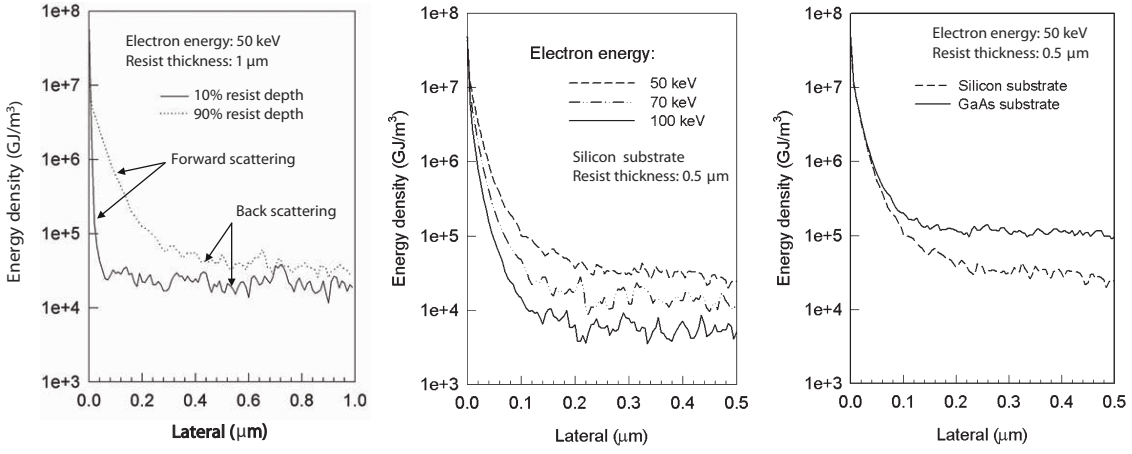


Figure 3.5: Point Spread Function (PSF) comparison for different (a) thickness of resist, (b) incident e-beam energy and (c) substrate material. Adapted from [122].

The non-uniform, radial distribution of exposure due to back scattering gives rise to pattern distortions and is given the name proximity effect. Pattern definition is affected by two types of proximity effects, intraproximity and interproximity, as shown in Figure 3.6. Intraproximity effects result in nonuniform exposure within a single pattern element, taking into account both the shape and size of the pattern. This leads to the pattern centroid receiving a higher exposure than the edges [125, 126]. For instance, the exposure received at the centre of a square pattern is twice that at the edge and four times that at the

corner [126]. In contrast, the interproximity effect gives rise to the nonuniform exposure between patterns. The proximity effect is enhanced between neighbouring patterns due to the scattering between these patterns. Therefore, geometries (e.g. linewidth) of individual patterns coupled with the distribution (spacing and size) of the neighbouring patterns have a considerable influence on the interproximity effect. Large, densely packed patterns are generally more exposed than small, isolated patterns. Hence, patterns that are densely packed together may be overexposed and even merge together [125, 126]. For definition of high resolution, high density patterns, both intraproximity and interproximity effects should be minimised during lithography in order to obtain well defined and well distributed features.

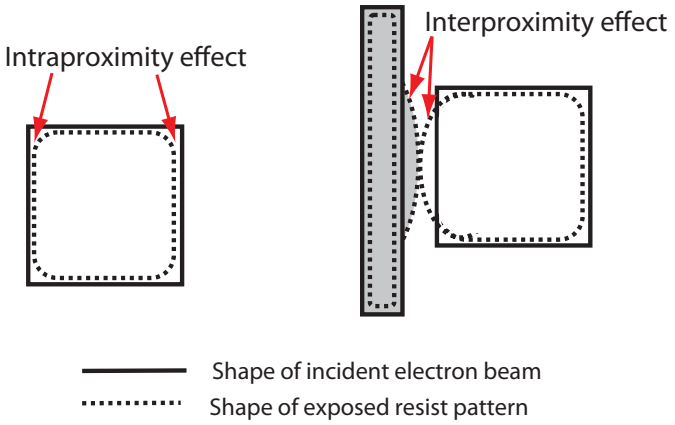


Figure 3.6: Schematic illustrating pattern distortions due to intraproximity and interproximity effects.

A number of corrective measures can be adopted to minimise the proximity effect. From a design perspective, optimising the distribution and packing density of patterns to be written reduces the effect. However, design constraints arising from device geometries may hamper the optimisation. In developmental MOSFET devices for instance, fairly large source and drain contacts need to be placed in close proximity. The proximity effect is also influenced by the aforementioned process conditions and can be minimised through the use of thin resists, high energy e-beam and low density substrates. Insertion of a thick, low density interlayer between the imaging resist layer and the substrate is another means of proximity effect minimisation. Multilayer resist techniques are often employed for this purpose. More commonly, proximity error correction software is used alongside the other corrective measures. This software exploits techniques that enable the exposure doses and the shape dimensions to be modulated during pattern writing, thereby compensating for the proximity effect [125]. The CATS package from Synopsys, that was used for fracturing all pattern files used in this work, comes with such software built-in.

3.2.3.3 Resist and Development

The choice of resist is important in EBL as it has a strong bearing on pattern definition. In general, sensitivity and contrast are the two most important properties of EBL resists, due to their direct relation to the achievable resolution. Other properties however may become important depending on the processing requirement, for example etch resistance is vital to subtractive patterning.

Resist properties are dependent partly on the resist chemistry, and partly on the processing conditions. The performance of a resist in a process is influenced by many parameters: exposure dose, e-beam energy, developer, development time and temperature, resist material and thickness, substrate material and pattern density [122, 127]. Many of these parameters are interlinked, and as such interact with the resist in a complex manner during lithography.

A given resist can be made more or less sensitive by modifying the process conditions. These conditions influence both forward scattering and back scattering, as discussed in Section 3.2.3.2. Higher e-beam energy, thinner resists, lower density substrates and sparsely packed patterns all contribute towards the lowering of sensitivity [122]. The sensitivity of a resist is further affected by its molecular weight. More chain scission events, leading to a larger exposure, are needed to achieve solubility in a resist as its molecular weight increases. As a result higher molecular weight resists have a lower sensitivity.

Development is an essential part of the pattern definition process as it affects the shape and size of patterned areas. Resist dissolution, during development, can be characterised as a diffusion-like process [127], which is influenced by the developer, temperature and time, resist molecular weight and exposure dose. For instance, in a positive tone resist an optimum developer would have minimum effect in the unexposed region whereas the exposed region would be effected maximally. This then increases the contrast between the two regions, and also provides for high resolution. Process contrast can change with both the choice and strength of developers [128]. Higher temperatures enhance the diffusivity of the resist polymers in a developer, leading to a faster rate of resist dissolution, hence process sensitivity is increased. In contrast, resists of higher molecular weight are less diffusive in a developer, which results in slower resist dissolution. Dissolution is also slower for smaller exposure doses. A slower dissolution rate translates to a longer development time, the result of which is reduced process sensitivity [127].

Contrast curves are used to characterise the sensitivity and contrast of a resist in a given process. These curves are obtained by performing a dose test, which is a prerequisite for defining and optimising new patterns. The test comprises of the exposure and

development of a series of identical patterns, in which each pattern is exposed at a different dose in a given experimental setup. The residual resist thickness at each dose is then measured and plotted as shown in Figure 3.7. The slope of the curve defines the resist contrast and is expressed as:

$$\gamma = \frac{1}{\ln(\frac{E_2}{E_1})}, \tag{3.5}$$

where E_2 is the minimum exposure dose required to fully develop out the resist and E_1 is the dose that marks the start of resist exposure [122]. A steeper slope corresponds to a higher resist contrast, while a smaller value of E_2 is indicative of higher resist sensitivity. For high resolution lithography, a resist with high contrast and low sensitivity is highly desirable. In addition, high contrast results in resist profiles having high aspect ratios and vertical sidewalls. The former profile makes possible the lithography of densely packed patterns while the latter facilitates subtractive patterning [122].

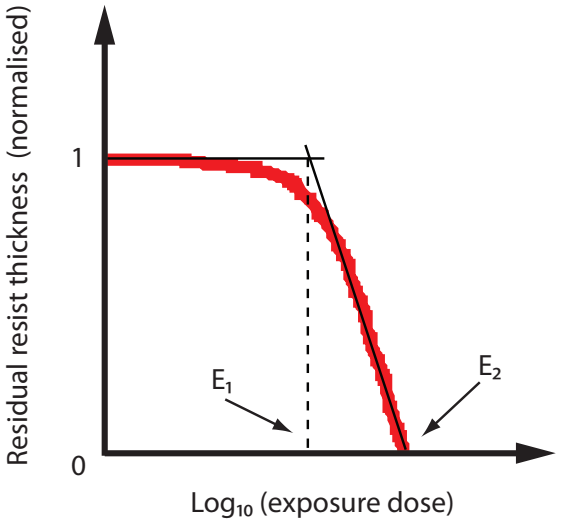


Figure 3.7: Contrast curve plot illustrating the residual resist thickness after development (normalised) as a function of the logarithmic exposure dose for a given experimental setup.

A positive tone resist that features heavily in EBL is poly-methyl methacrylate (PMMA). It offers high resolution patterning capability which in turn has enabled the realisation of features as small as 5 nm. PMMA is a low sensitive resist which comes in a number of different molecular weights, with sensitivity increasing slightly with decreasing weight [122]. The most commonly used developer for PMMA is methyl isobutyl ketone (MIBK), a strong solvent, diluted with isopropyl alcohol (IPA), a non-solvent, to differing concentrations. Weaker MIBK/IPA developers provide for higher contrast and higher resolution, as shown in Table 3.2. Another effective developer for PMMA is water/IPA.

Table 3.2: Comparison between different developer concentrations and their corresponding effect on resolution. Adapted from [122].

MIBK:IPA concentration	Contrast	Sensitivity	Resolution
1:3	Very high	Low	Extremely high
1:2	High	Medium	Very high
1:1	Medium	High	High
Pure MIBK	Low	Very high	Low

The combination of the two non-solvents results in what is called cosolvent, which has a solvent strength larger than its parts, and provides for high sensitivity and high contrast simultaneously [128].

The differential sensitivities between the different molecular weights, makes PMMA highly desirable for use in a multilayer resist process such as the one detailed in Section 3.3.2. One of the main drawbacks of PMMA however is its poor etch resistance [117], which makes it unattractive for subtractive processing. Hydrogen silsesquioxane (HSQ) on the other hand has high etch resistance, making it ideal for this purpose. This is a negative tone resist that has a similar contrast and sensitivity as PMMA, which allows for high resolution pattering into the sub-10 nm range [122].

3.2.3.4 Alignment

Pattern overlays are often required to realise a fully functional device. This entails two or more levels of lithography for which the patterns defined in each of these levels need to be accurately positioned with respect to one another. This process of matching patterns from one lithography level to the next is referred to as alignment.

Alignment is essential to correct for the various errors that arise during pattern overlay. The potential sources of error are depicted in Figure 3.8 with the aid of a transistor fabricated using two levels of lithography in a gate-last process flow. A translational error results in the gate being offset from its desired position in the x-axis, y-axis or both. Angular displacements about the gate centre, due to a change in x-axis as a function of y-axis or vice versa, give form to rotational errors. A gain error causes the gate to be scaled either in one axis as a function of that axis or in both axes as a function of the corresponding axes, leading to oversized or undersized gate definition. Keystone errors result from changes in one axis as a function of both axes. These misalignments can arise from a number of different sources, for example substrate loading errors, temperature changes within the tool or positional drift of the beam with time. For instance a 0.1 °

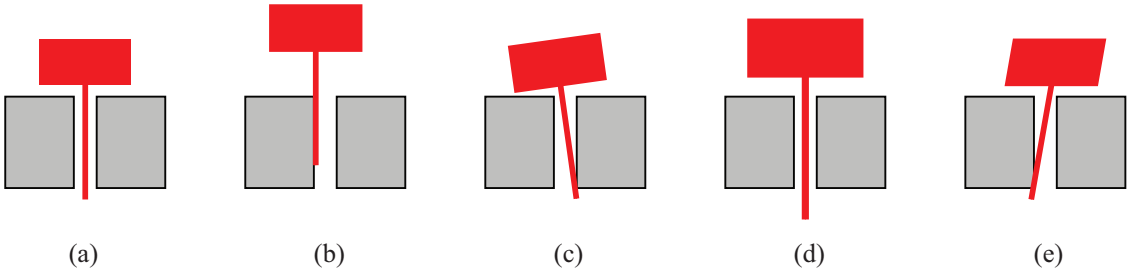


Figure 3.8: Schematic showing alignment errors that can arise during gate pattern overlay to a previously defined source/drain pattern layer for transistor fabrication: (a) correct alignment, (b) translational error, (c) rotational error, (d) gain error and (e) keystone error.

rotation of a 1 mm^2 sample about its centre amounts to a positional error of approximately 2 μm, which in most fabrication cannot be tolerated.

Pattern overlay errors are corrected by aligning the pattern layout to be written to pre-existing reference marks on the substrate, referred to as alignment markers. These markers are written as the first lithographic level, either on their own or often married to the first pattern layout as part of a device process flow, to save on a lithography step. This enables subsequent lithography levels to utilise markers defined in the first level for alignment. This is known as zero-level alignment and is preferred over level-to-level alignment. In the latter technique, subsequent pattern overlays are aligned using the markers from the most recent overlay. As a result, the overall alignment error is compounded by the alignment errors introduced from each pattern overlay.

The attributes of the alignment marker, the marker detection methodology, as well as alignment strategies used can significantly affect pattern overlay accuracy. These are considered in the following paragraphs.

The shape and physical structure of the markers is of importance to alignment. Square markers tend to be commonly used, although other more complex geometric shapes can be employed. The markers are defined in two ways; the first is by etching into the substrate to form pits or raised mesas (topographical marker), the other is to deposit materials of a higher atomic number, relative to the substrate, to leave a raised section (z-contrast marker). These marker types are shown in Figure 3.9. A solid-state back scatter electron (BSE) detector is used to detect the back scattered electron signal as a function of beam position as the e-beam scans across the markers. These back scattered electrons are a result of the electron-substrate interactions that occur from e-beam exposure. Materials of a higher atomic number produce more back scattered electrons (Section 3.2.3.2). As a result when the beam scans across from the substrate to the marker, the electron signal

level changes according to the amount of back scattering detected [129]. To achieve signals of high contrast, markers are generally made of metals such as gold (Au) which have much higher atomic numbers compared to the substrate, e.g. gallium arsenide (GaAs). On the other hand, topographical marker detection is based on a different mechanism. The step or contour of the marker creates a modulation of the back scattered signal. This modulated signal is composed of a slow and fast transient. The slow transient is associated with the beam approaching or receding from the step, whereas the fast transient is related to the steep crossing of the step [129]. These two effects combined produce a signal as illustrated in Figure 3.9, which comprises two spikes corresponding to the points of topographical change. The quality and contrast of the signal is further affected by marker height. The thickness of z-contrast markers needs to be about $50-100\,\text{nm}$, while topographical markers are usually microns deep.

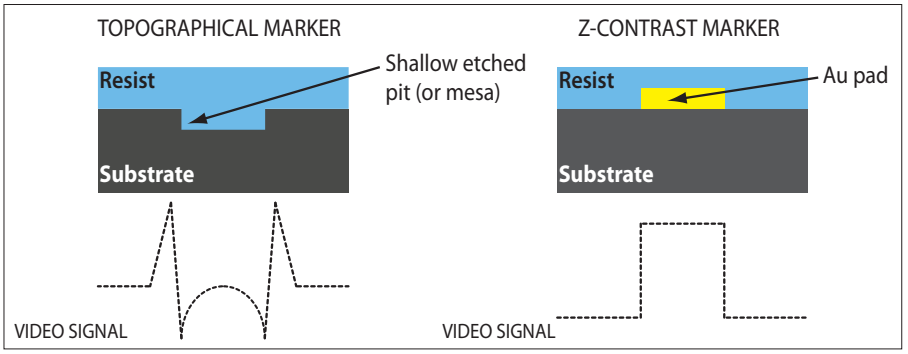


Figure 3.9: Alignment mark types and their corresponding back scatter profiles.

Marker detection is based on a search and locate algorithm known as "mark locate", which is conducted automatically in software. During the search, marker edges are scanned back and forth by the beam several times. The average of these scans is used to locate the edges, and corresponding centre of the marker. Four markers are typically required to correct for all of the aforementioned pattern overlay errors and are best placed at the corners of a layout. This configuration, referred to as "global alignment", marks out the grid boundaries within which the pattern layout should be placed for optimum alignment as shown in Figure 3.10. Nonetheless, alignment can still be successfully obtained using fewer markers. However, with decreasing number of markers, the sensitivity of the alignment process to specific overlay errors is lost. For instance, keystone errors are not accounted for with only three markers.

For large pattern layouts, the separation between the markers increases their positional uncertainties. This gives rise to local variations in alignment of patterns across a large layout. A two-stage alignment strategy is adopted to circumvent this effect. In the first

instance, a "global" alignment is performed as normal. This is followed by "cell" alignment of smaller areas of patterns. In order to implement this, layouts are divided up into pattern arrays, an example of a 2×2 array is shown in Figure 3.10. Each pattern in the array has its own set of "cell" alignment markers, with a smaller marker grid providing for a more accurate determination of marker positions. However, this comes at the expense of throughput as the time required for alignment increases.

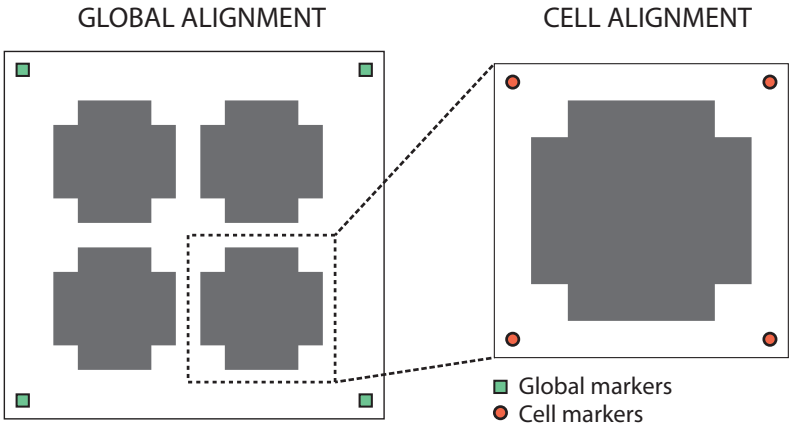


Figure 3.10: Schematic of "global" and "cell" alignment procedures during pattern overlay.

The overlay accuracy that can be achieved is inherently limited by the ability of the search algorithm to locate marker edges precisely, which in turn requires markers to be well-defined. Square markers are capable of alignment accuracies in the order of 15 nm. However, this is based on the assumption that the quality of the markers edges is near-perfect. In reality, this number is often bigger. One of the process modules developed in this work, described in Chapter 6, requires overlay accuracies of a few nm for which better alignment techniques are needed. For this the correlation-based alignment technique developed in [130], with a measured overlay accuracy of 0.63 nm, was adopted. In this technique Penrose tilings are used as markers, as shown in Figure 3.11. The increased number of marker edges provides for a more accurate determination of the marker position. Moreover, its compatibility with well-established fabrication techniques, such as lift-off, makes it highly suited for implementation.

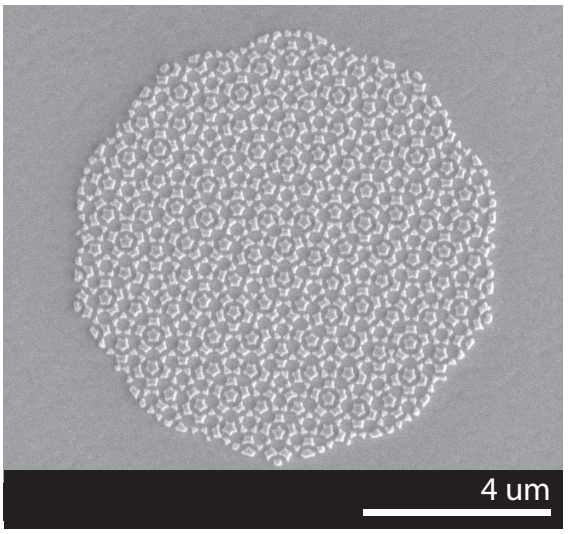


Figure 3.11: SEM micrograph of Penrose tilings defined as a z-contrast marker via e-beam evaporation of Al(20 nm)/Au(100 nm) and lift-off.

3.3 Pattern Transfer

Pattern transfer, using the latent resist image as a stencil, can be either additive or subtractive. Material is added to the exposed resist regions in the additive process, whereas in the subtractive transfer material is removed from the exposed regions. The two most commonly used pattern transfer processes are lift-off (additive) and etching (subtractive). In the following paragraphs, these processes are described in the context of metal feature definition, since this was the most widely used form of patterning in this work.

3.3.1 Subtractive Patterning

In subtractive processing a metal film is first blanket deposited over the entire substrate and then selectively masked with negative resist. Thereafter, a suitable etch process is used to remove the metal film in the unmasked, resist-free regions thereby defining the metal feature. The associated etchant can be in the form of a chemical solution (wet-etching) or gas/plasma process (dry-etching) described in Section 3.5.

This technique, in general, is suitable for substrates that are highly resistant to etchant damage, for instance plasma-induced damage from dry etching, and non-composite metal films. Key considerations for usage include the resistance of masking layers to metal etchants, availability of appropriate metal etchants and deposition techniques that provide conformal coverage. While physical vapour deposition (PVD) techniques are usually

used for metal deposition, chemical vapour deposition (CVD) techniques, described in Section 3.4.2, are preferred for device topographies featuring high aspect ratios.

Subtractive processing, due to its high reliability and yield, commonly features in silicon IC fabrication for patterning refractory metals in conjunction with dry etch techniques. For instance, interconnect technology traditionally used subtractive aluminium prior to the incorporation of copper [131]. On the other hand, additive processes are more common in III-V processing. This arises from the fact that the dry-etch induced damage cannot be cured by post-etch annealing in III-Vs, unlike the case in Si [14]. Associated damage can manifest in the form of physical damage and/or degradation of lattice properties. The former could cause surface roughening while the latter could induce surface states [132].

3.3.2 Additive Patterning

Lift-off is a standard technique used in III-V processing to define metal features. For lift-off to be effective a break in the deposited metal film is required. This is attained using a "line-of-sight" metal deposition technique in parallel with a multilayer resist process that provides an undercut (or overhung) resist edge profile. Accordingly, metal deposition via evaporation and a bilayer resist [133] technique were used in lift-off processing of patterns defined by EBL in this work. The process flow incorporating these techniques is illustrated in Figure 3.12.

In the bilayer resist process two layers of PMMA resist, with differing molecular weight, are spin-coated on the substrate and exposed, with the higher molecular weight resist as the top layer (Figure 3.12(a)). Given the higher sensitivity of the lower molecular weight resist, the solubility of the bottom resist is greater than the top layer. Therefore, during development the bottom resist develops out faster laterally compared to the top resist, resulting in a undercut resist edge profile (Figure 3.12(b)). The subsequent metal

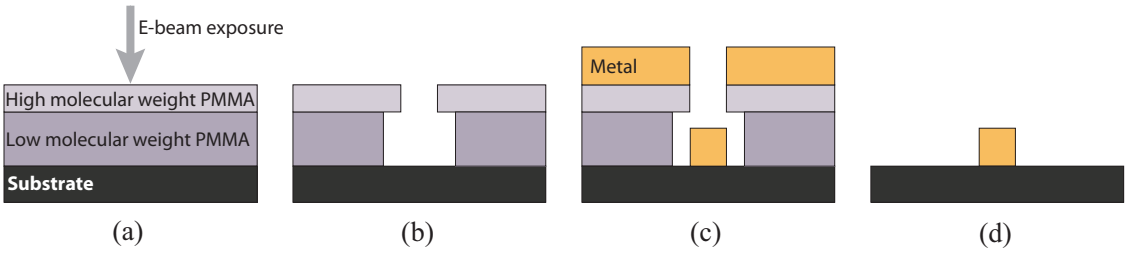


Figure 3.12: Process flow for lift-off comprising of (a) resist application and e-beam exposure, (b) resist development, (c) metal evaporation and (d) lift-off.

evaporation creates a discontinuous metal film over the resist edge of the exposed regions such that the resist sidewalls are metal-free (Figure 3.12(c)). In the final step the use of a solvent, such as acetone, acts through the exposed sidewalls to remove the resist and "lift-off" the overlying metal in the unexposed regions, with only the patterned region retaining metal (Figure 3.12(d)).

It is often the case that device features from previous processing are present on the substrate prior to metal deposition and lift-off. Therefore, a relatively thick bottom resist layer, in addition to providing a good undercut resist profile, is used to planarize the surface topography. Moreover, for effective lift-off this layer needs to be at least as thick as the deposited metal. In contrast, the top resist layer is usually relatively thin in order to achieve high resolution pattern definition [133].

Lift-off is a simple and cheap process. The mild, less damaging processing involved coupled with composite metal stack compatibility make lift-off particulary favourable for III-V processing. However, depending on the material used, yield, retention and flagging issues can crop up [134].

3.4 Thin Film Deposition

Thin films serve a variety of functions for the many processes involved during the fabrication of CMOS ICs. The active area of devices are defined by growth of thin semiconductor films, insulator films are used for gate dielectrics, isolation, passivation, implantation masks, diffusion barriers and sidewall spacers. Similarly interconnects, vias, contacts and bondpads all require thin films of conductors such as aluminium, copper and tungsten [48]. The relevant choice of deposition technology is often guided by the process-specific requirements of the deposited film material as needed at different stages of device fabrication. In this section the techniques used for epitaxial materials growth, dielectric deposition and metal contact formation, as relevant to fabricated devices, are presented.

3.4.1 Epitaxial Material Growth

Epitaxial growth, in which a grown layer adopts the same lattice constant and crystal symmetry as the underlying material, has enabled semiconductor device designs to be advanced. By taking advantage of this growth technique, bandgap engineered and quantum confined device structures have been realised, a prime example being the HEMT [135].

The operational success of the HEMT hinges on the precise control of the device layers down to the monolayer during growth. In particular, the ability to achieve atomically

abrupt, smooth compositional transitions and planar doping profiles are critical [135]. Similar rules apply to the implant-free III-V p-MOSFETs developed in this work as the device concept, highlighted in Chapter 5, is based around the HEMT with the addition of an oxide layer serving as a gate dielectric.

Among the various epitaxial growth techniques available, MBE and MOCVD are the two most widespread. Of these, MBE is particularly suited for high levels of interface control and precision [135]. In line with this, the growth of III-V active layers for this work was performed using MBE.

3.4.1.1 Molecular Beam Epitaxy

In the MBE growth process the elements required for the epitaxial layer are co-evaporated from sources in the form of molecular beams. These beams impinge onto a heated crystalline substrate, where they chemically react to produce film growth. Although MBE growth appears conceptually simple, there are a number of working parts to the process, these are described in the following paragraphs.

The basic components of a typical MBE system are the loadlock, buffer and growth chambers, shown in Figure 3.13. The entire system is held under ultra-high vacuum (UHV) and the chambers are interconnected through gate valves, enabling the transfer of substrate wafers between them. Substrate wafers are introduced and removed via the loadlock, without affecting the vacuum conditions held in the other chambers. The buffer chamber is used as intermediary, between the loadlock and growth chamber, for sample preparation and de-gassing. In-situ diagnostic tools, needed for surface analysis of the substrate prior to and after film growth, are often housed in this chamber, e.g. x-ray photoemission spectroscopy (XPS). The growth chamber comprises the effusion cells,

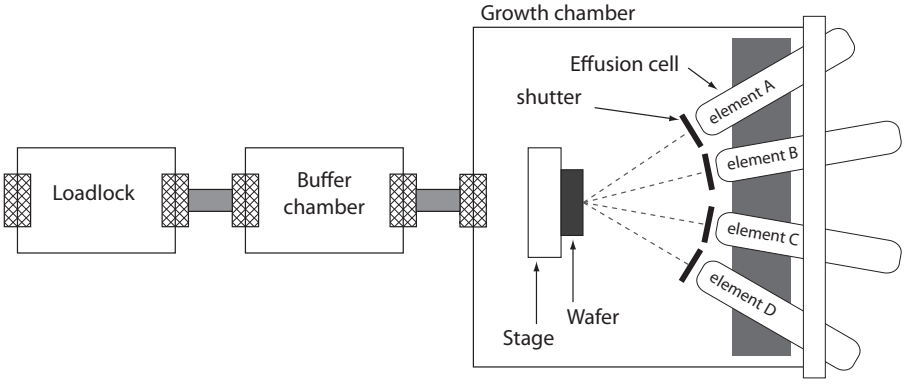


Figure 3.13: Schematic of molecular beam epitaxy (MBE) system.

mechanical shutters and stage. High-purity elemental sources (evaporants) are contained within the effusion cells. The evaporant flux impinging on the substrate is determined by the temperature the effusion cell is heated to. Mechanical shutters located at the front of sources enable the flux to be interrupted during growth. The substrate is placed on the stage, which provides heating, rotation and translation capabilities. During growth, the substrate is normally heated, which typically for GaAs is 400 - 600°C. The growth chamber is maintained at UHV conditions of 10^{-10} - 10^{-11} Torr for high quality film growth [135].

Thin films of precise composition, thickness and doping profile are realised by fine tuning the growth parameters in real-time. MBE growth chambers usually come with a number of in-situ diagnostic tools for monitoring the growth process, of which reflection high energy electron diffraction (RHEED) is the most common. In this technique an e-beam is directed onto the growing surface at small grazing angles, which is then Bragg diffracted and imaged onto a fluorescent screen. If the surface is perfectly planar and ordered, the diffraction pattern is a set of rods, separated by the spacing of surface features. In reality though beam dispersion and divergence, coupled with lattice imperfections, result in the surface not being perfectly planar. As a result the diffracted image appears as a set of streaks with modulated intensities. In the case of the surface being rough, the image consists of spotty patterns. Apart from surface flatness, surface reconstructions are also identified from the RHEED patterns. This provides information on the topography and composition of the growing layers. Additionally, growth rates are extracted from the intensity oscillations observed in the RHEED patterns [135]. RHEED is therefore an invaluable tool in monitoring and optimising growth conditions, enabling films of desired thickness and composition to be obtained.

Many of the features and process conditions characteristic of MBE are advantageous to film growth, as shown in Table 3.3. However, there are a few issues to contend with when using the tool. For high levels of interface control and precision a slow growth rate and UHV environment are vital. However, this comes at the expense of throughput and costs of operating the tool. The top up of source material requires the chamber to be exposed to air (vented), which compromises the UHV conditions. To get the system back up to its operational condition, a high temperature bake is usually performed for an extended period of time, which adds to the complexity of the tool. At the University of Glasgow a 24-hour bake and pump-down is typically required for the MBE tool to be operationally ready following chamber venting.

Table 3.3: MBE process features, and the resulting effects and benefits to film growth [135].

Process feature	Effect	Benefit
UHV (10^{-10} - 10^{-11} Torr)	Negligible beam scattering	Uniform compositional profile
	Impurities impinging on substrate is minimised	High purity film growth
Source material purity (99.99999%)	Impurity levels reduced	High purity film growth
Slow growth rate (1 μm/hr)	Well-controlled growth	Accurate thickness of grown films
Reduced growth temperature (e.g. 400 - 600°C for GaAs)	Sufficient surface diffusion	Layer-by-layer growth
	High interface control	Abrupt, smooth compositional & doping profile
	Minimal bulk diffusion	Compositional & doping profiles unperturbed
Fast shutter response (0.1 s)	Abrupt growth interruptions	Monolayer abruptness of composition
	Well-controlled growth rate	Accurate thickness of grown films
Substrate Rotation	Lateral compositional & growth rate variations reduced	Large-area growth uniformity
Sufficiently large separation between source & substrate	Source angular separation minimised - similar flux distributions	Uniform compositional profile
In-situ diagnostics (e.g. RHEED)	Real-time analysis of growth: thickness, composition & surface quality	Growth rate & compositional control

3.4.2 Dielectric Deposition

Dielectric thin films that are highly conformal and that provide good step coverage are largely obtained using chemically-based processes such as CVD in industry. In CMOS, silicon nitride (Si_3N_4) films deposited by means of plasma-enhanced chemical vapour deposition (PECVD) are routinely used for device passivation [48]. Although gate dielectrics have been perennially grown by thermal oxidation, the transition to hafnium-based high-k dielectric in CMOS has led to the adoption of atomic layer deposition (ALD) [10]. Commensurate with this, ALD has become the standard for high-k gate dielectric deposition in research centred around MOSFETs, not just limited to Si. The devices presented in this work feature ALD aluminium oxide (Al_2O_3) gate dielectrics, while Si_3N_4 based on inductively-coupled plasma chemical vapour deposition (ICP-CVD) was briefly investigated as an encapsulation layer.

3.4.2.1 Inductively-Coupled Plasma Chemical Vapour Deposition

In a CVD process, chemical reactions involving the components of the vapour phase at or near the substrate surface result in film formation. The constituents of the deposited films

are derived from the precursor gases that make up the vapour. In general, the process encompasses three steps [136]:

1. The precursor gases traverse to the substrate surface.
2. At the surface the gases undergo adsorption followed by dissociation into reactants, which then migrate to reaction sites where chemical reactions ensue.
3. By-products of these reactions are desorbed and removed from the surface.

Typically, Si_3N_4 films are deposited by means of low-pressure chemical vapour deposition (LPCVD) or plasma-enhanced chemical vapour deposition (PECVD). The former technique is performed at elevated temperatures ($> 600°C$) [137] which provide the requisite energy needed to drive the chemical reactions during deposition. Such high thermal budgets though are undesirable in III-V process flows. Neither epitaxially grown III-V thin film layers, nor subsequent process features such as Schottky or ohmic contacts, can survive such high temperatures without detriment [138]. PECVD, on the other hand, makes use of a plasma as a source of non-thermal energy to drive the chemical reactions, thereby enabling the deposition to occur at relatively lower temperatures ($< 400°C$) [136, 138]. In the plasma, the precursor gases interact kinetically with the high-energy electrons. This results in an ionisation of the gases such that the plasma is filled with chemically reactive species (free electrons, energetic ions, free radicals etc.). These active species are then used for film formation employing the aforementioned CVD process steps.

PECVD is capable of depositing Si_3N_4 films at 300°C [137]. However, obtaining stoichiometric Si_3N_4 composition is difficult. This in turn affects the electrical, mechanical and optical properties of the films [136–138]. Nonetheless, this may not be a serious issue if the as-deposited film serves its intended purpose, regardless of the resultant film properties. A more pressing concern is plasma-induced damage to sensitive surface layers. Devices incorporating thin layers of III-V semiconductors and gate dielectrics are susceptible to such damage, which could lead to performance degradation.

Plasma-induced damage arising from PECVD can be substantially minimised through the use of inductively-coupled plasma chemical vapour deposition (ICP-CVD). This technique presents with a means of obtaining high quality, high-density, low stress Si_3N_4 films at considerably lower deposition temperatures. At the University of Glasgow, a room temperature ICP-CVD technology based on an Oxford Instruments PlasmaLab System 100ICP180 has been developed for Si_3N_4 deposition [137]. This is schematically depicted in Figure 3.14. The system comprises of a discharge chamber where the plasma is formed

and a process chamber where the deposition occurs. A radio frequency (RF) generator connected to the ICP coils sets up a potential within the discharge chamber which in turn produces the plasma. The plasma dissociation and density of incident ions is controlled by the ICP coil power. A second RF generator is used to separately power the chuck, thereby enabling independent control of the energy of the ions incident on the substrate. Applying high power to the ICP coils with the chuck held at zero bias produces a high-density plasma, this enables the process temperature to be lowered. In addition plasma-induced damage to the substrate as well as the stress levels of deposited films are minimised.

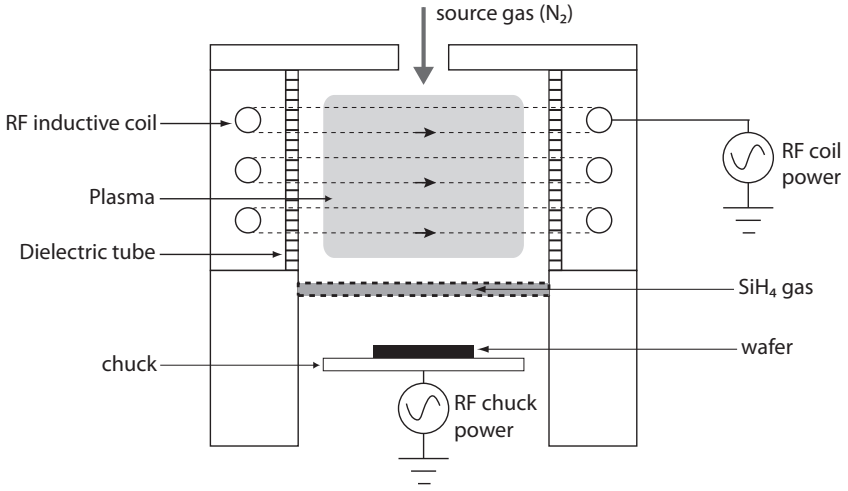


Figure 3.14: Schematic of inductively coupled plasma chemical vapour deposition (ICP-CVD) system. Adapted from [137].

Silane (SiH_4) and either gaseous nitrogen (N_2) or ammonia (NH_3) are used as precursors to form the constituents of the Si_3N_4 film. PECVD uses NH_3, since N_2 has a lower dissociation efficiency. As a result hydrogen is incorporated to a certain degree in the deposited Si_3N_4, thereby lowering the quality and purity of the film. On the other hand, N_2 is used in ICP-CVD as the high-density plasma source provides for more efficient ion dissociation, which is almost an order of magnitude higher than in PECVD [139], hence improved film quality is obtained.

3.4.2.2 Atomic Layer Deposition

While ALD is a subclass of CVD, the process of film growth distinguishes it from its counterparts. In CVD precursors are injected into the growth chamber simultaneously and continually, such that surface reactions leading to film growth ensue in a somewhat uninhibited fashion. In contrast film growth by means of ALD occurs in a cyclic manner, in which one growth cycle comprises four sequential steps [140]:

1. A reaction of the first precursor that is pulsed into the reaction chamber.
2. A purge of the chamber to remove the excess gases.
3. A reaction of the second precursor that is pulsed into the chamber.
4. A purge to remove excess gases.

The well-established Al_2O_3 deposition on a substrate surface containing hydroxyl (OH) groups, using the precursors of trimethyl-aluminum (TMA) and water (H_2O), can be drawn on to demonstrate the stepwise ALD growth process. Prior to deposition, the substrate surface is stabilised to a controlled state usually via an in-situ heat treatment (50-400°C). This is followed by the injection of gas-phase TMA ($Al(CH_3)_3$) molecules into the chamber, which then through surface reactions are chemisorbed. TMA reacts only with the OH ligands on the surface, the result of which is the bonding of aluminium ions (with two methyl (CH_3) ligands attached) to the oxygen, together with the release of methane (CH_4) by-products [140] (Figure 3.15(a)). This is the 1st half-reaction of the ALD cycle and is described as:

$$AlOH^* \; + \; Al(CH_3)_3 \; \longrightarrow \; AlOAl(CH_3)_2^* \; + \; CH_4\,, \tag{3.6}$$

where the asterisk denotes the species residing on the surface [141, 142]. This reaction continues to the point of saturation, meaning all available surface sites have undergone reactant chemisorption such that no further adsorption is possible. Thus, the reaction ceases when the surface is saturated. This mechanism of film growth, referred to as self-limiting [141, 142], is characteristic of ALD, which results in the deposited film being of a constant thickness in each half-cycle. Prior to introducing the second precursor, the chamber undergoes a purge. Excess gases in the form of un-reacted precursor molecules and reaction by-products are evacuated, leaving the surface CH_3-terminated (Figure 3.15(b)). This completes the 1st half-cycle. The 2nd half-cycle is initiated by introducing H_2O precursor molecules into the chamber, which then react with the dangling CH_3 ligands on the surface. As a result aluminium-oxygen (Al-O) bridges are formed and CH_3 ligands are replaced by OH groups, with CH_4 produced as by-products [140] (Figure 3.15(c)). This is the 2nd half-reaction and is described as [141, 142]:

$$Al(CH_3)^* \; + \; H_2O \; \longrightarrow \; AlOH^* \; + \; CH_4\,. \tag{3.7}$$

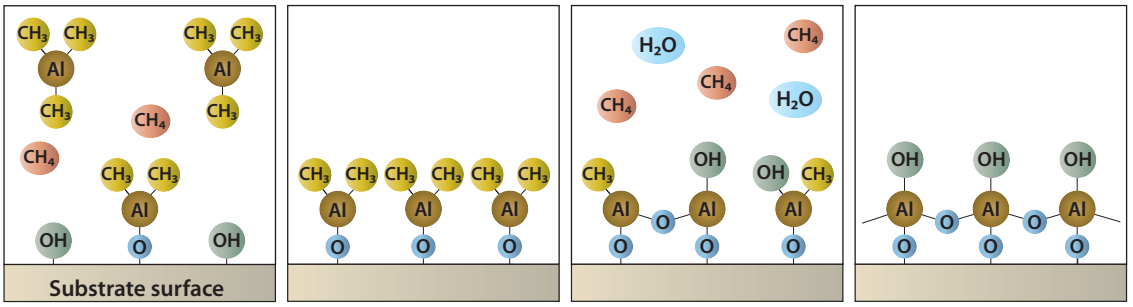


Figure 3.15: An ALD growth cycle schematically depicting Al_2O_3 deposition from TMA ($Al(CH_3)_3$) and H_2O precursors on an OH-terminated substrate surface: (a) the pulsed TMA precursor is chemisorbed until the surface is saturated. Al ions bond with O, and in the process produce CH_4 as reaction by-products, (b) un-reacted TMA and CH_4 by-products are purged, leaving behind a CH_3-terminated surface, (c) pulsed H_2O precursor is chemisorbed until the surface is saturated. Al-O bridges are formed and OH groups replace CH_3 ligands, with CH_4 generated as by-products and (d) un-reacted H_2O and CH_4 by-products are purged, leaving behind an OH-terminated surface.

As with the 1st half-cycle, a purge is carried out in the chamber to remove all remnant gases, leaving the surface OH-terminated [140] (Figure 3.15(d)). With this, the 2nd half-cycle is complete. This also concludes one full ALD growth cycle with one monolayer of Al_2O_3 deposited on the substrate. During deposition this cycle is repeated as many times as needed to achieve the desired film thickness.

The cycle time varies between 0.5 to a few seconds depending upon the aggressiveness of the growth reactions. The film growth rate per cycle also varies between 0.01–3 nm. This variation stems from two factors; the size of the precursor molecules and the number of surface sites available for adsorption. Larger precursor molecules, due to steric hinderance between the molecules, and a smaller number of adsorption sites tend to limit the number of molecules that can be surface-adsorbed [141].

The ALD growth process has several advantages, the deposited film thickness is dependent only on the number of cycles, hence the thickness can be controlled precisely and in a simple manner, down to the atomic scale, which is particularly important for the growth of ultra-thin films [141, 142]. Additionally, the self-limiting growth removes the importance of precursor flux homogeneity to the process [142]. Thus, the deposition is smooth, highly conformal and provides excellent step coverage on substrates featuring complex topography and high-aspect ratio features. Moreover, film growth is highly reproducible and provides for large area uniformity [141, 142]. The filling of all available surface adsorption sites also produces films that are continuous and pinhole-free [142],

features critical for CMOS gate dielectrics. Furthermore, the growth occurs at relatively low process temperatures ($< 400°C$), making it compatible with III-V processing.

3.4.3 Metal Deposition

Metallisation by means of ALD is gaining popularity for MOS devices. The complete growth of the gate stack in-situ provides for a gated oxide region that is shielded from further processing and its associated damage. However, the patterning of the gate entails subtractive processing, which for III-V materials is damaging. Traditionally metals are deposited using physical vapour deposition (PVD) techniques such as evaporation and sputtering, which are line-of-sight techniques. The better step coverage from sputtering may cause problems during lift-off, which is the technique adopted for metal definition. Furthermore, sputtering can induce damage to III-V substrates. For these reasons evaporation was used for metallisation in this work.

3.4.3.1 Evaporation Techniques

Evaporation is a well established metal deposition technique which occurs in three stages:

1. The source material is heated to its melting point such that it evaporates.
2. The evaporant is transported from the source to the substrate.
3. On the substrate the vapour species condense, and nucleate leading to film growth of the source material.

In evaporation processes different methods can be used for heating the source. Resistive and e-beam heating tend to be the two most commonly used methods. In resistive evaporation, commonly referred to as thermal evaporation, a high current is passed through a crucible made of refractory metals, such as a tungsten boat, which contains the source material. In contrast, a high-intensity electron gun is used in e-beam evaporation to focus an electron beam onto the surface of the source material, resulting in its vaporisation.

The evaporator tools used for the work of this project are shown schematically in Figure 3.16. The e-beam evaporator has two shutters while the thermal evaporator has only one. The additional source shutter in the e-beam evaporator provides a means of stabilising the evaporation rate and its corresponding flux prior to the shutter being opened. The substrate shutter is only opened when the desired evaporation rate is reached.

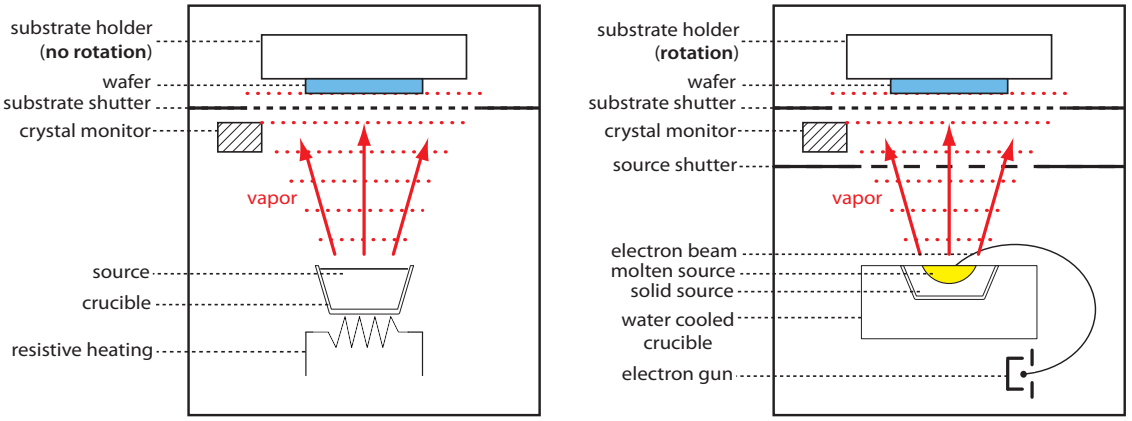


Figure 3.16: Schematic of (a) a resistive evaporator and (b) an electron beam evaporator.

This two-shutter method protects the substrate from spikes in the evaporation rate that may occur during the initial heating process.

The source material is situated in a water cooled copper hearth in the e-beam evaporator. In this manner, the source forms its own crucible during the localised heating process, with only the centre portion of the source melting. As a result source contamination is minimised in e-beam evaporators as opposed to thermal evaporators.

Evaporation is performed under high-vacuum conditions. Under these conditions the mean free path (MFP) of the evaporants is larger than the chamber size. As a result a line-of-sight path is taken by the atoms during their transport from the source to the substrate. While the e-beam evaporator could reach evaporation pressures of less than 10^{-6} Torr consistently, pressure in the thermal evaporator is in the range of $10^{-6} - 10^{-5}$ Torr.

The thickness of metal deposited is greatest below the direct line-of sight of the source and decreases away from it. To achieve a uniform thickness deposition across the entire length of the sample the e-beam evaporator uses a rotating substrate holder, whereas the holder is static in the thermal evaporator. Although e-beam evaporators are preferred for metal deposition for the aforementioned reasons, x-ray emission from the beam can cause substrate damage [143].

3.5 Thin Film Removal

Etching is a subtractive process that is important for many fabrication needs. It is an industry-standard technique that is used in a myriad of processes in CMOS. For instance,

chemical mechanical polishing (CMP) is used for planarization in a number of processes. In III-V technology etching is required for device isolation to define the mesa in FETs and gate recess etching of HEMTs.

There are two types of etching processes, wet and dry, that can be used for pattern transfer. Wet etching, as its name implies, uses chemical etchants in the form of solutions to remove material. Dry etching on the other hand is a plasma based technique that uses chemical reactions, physical reactions or a combination of the two, to remove material.

The main issues to consider when selecting a particular etch technique for a process are as follows:

1. *Etch rate* - the time required for a specified etch process to be completed.
2. *Etch Profile* - refers to the directionality and degree of isotropy of the etch process in removing material. Isotropic etching removes material in all directions while anisotropic etching proceeds unidirectionally.
3. *Critical dimension control* - refers to the smallest feature size that can be etched without being affected by the etch process.
4. *Selectivity* - the degree to which the etch can differentiate between the layer to be etched from the layer to be retained.
5. *Damage* - refers to an etch process causing detriment to semiconductor properties resulting in degraded device performance.

3.5.1 Wet Etching

In wet etching, a chemical process ensues between reactants in the etchant solution and the surface of the substrate. This process proceeds in three basic steps:

1. Reactants are transported to the substrate surface.
2. A chemical reaction occurs at the surface.
3. The resultant products are transported away from the surface.

The etching chemistry can be characterised as being either diffusion-limited or reaction-rate limited. If the rate at which the etching progresses is dependent on the transport of the reactants to and away from the surface, it is diffusion-limited. Otherwise, if the reaction step decides the etching rate, then it is reaction-rate limited. The

conditions under which the etching proceeds dictates the overriding etching chemistry. These are etch concentration, composition and temperature [144].

Reaction-rate limited etching is particularly affected by temperature changes, but is relatively insensitive to agitation. On the other hand diffusion-limited etching is largely affected by the degree of agitation, with the etch proceeding faster with increased agitation, while being relatively insensitive to temperature. Both etching chemistries are affected by concentration, with reactant-rate limited etching favoured at lower concentrations in contrast to diffusion-limited etching favoured at higher concentrations. Surface quality can be seriously degraded when etch rates are extremely quick. Moreover, such fast rates are hard to accurately control. Since wet etching is highly dependent on these conditions, achieving run-to-run reproducibility can be difficult.

In general wet etches are considered to be isotropic in nature. However, etch profiles resulting from wet etching are highly dependent on the crystallographic orientation of the substrate, which can lead to orientation-dependent anisotropy [144]. In the case of GaAs, alignment of the etch mask parallel to $\langle 011 \rangle$ direction results in two sides (along the same direction) of the mask to have etch profiles that are sloping while the orthogonal sides will have undercut profiles. This crystallographic dependence of the etch profile can have consequences, for instance when defining gate extensions along the mesa edge in FETs. The sloping profile will enable successful gate definition while the undercut profile will cause a discontinuity in the gate as it runs along the edge.

Another effect of wet etching is the often near equal rates at which vertical and lateral etch fronts proceed. The lateral over etch can become significant for deep vertical etches. This makes dimensional control of critical features difficult. As such, wet etching is often used for features that are larger than at least a few μm, with dry etching preferred for smaller features. Moreover, the chemical character of the surface of the substrate prior to etching needs to be taken into account. Very often native oxides form on surfaces from ambient exposure. This is more pronounced for surfaces containing aluminium or antimony. Depending on the sample history, initial etch rates can vary from one wafer to the next for the same material composition. Typically a native oxide etch precedes the material etch to ensure uniformity of etch rates from run-to-run. This native oxide etch is particularly important during the de-oxidation step prior to metal deposition. Any residual oxide after the etch can be detrimental to device performance. Ohmic contacts, for instance, may exhibit increased resistance which will degrade device performance.

There are a number of advantages to wet-etching, firstly compositional etch selectivity can be achieved for a wide variety of material compositions. For example, a solution

comprising of citric acid and hydrogen peroxide (H_2O_2) can be used to selectively etch InGaAs on InAlAs, as well as InGaAs on InP [145]. An important benefit of wet etching in III-V processing is the etched substrates are virtually damage-free, as opposed to dry etched substrates which suffer from damage. Moreover, wet etching is a simple process that can be easily implemented given that a wide range of chemical solutions exist for most materials.

3.5.2 Dry Etching

Dry etching comprises a number of different techniques that include, sputter etching, ion-beam etching and reactive ion etching. Of these reactive ion etching (RIE) is the most widely used technique for realising nanoscale features during pattern transfer and therefore is described in the following paragraphs.

Chemical and physical processes are combined to produce the mechanism leading to material etch in RIE and is explained with the aid of Figure 3.17. The application of RF (13.56 MHz) power to the bottom of the two parallel plates, with the other grounded, dissociates the gas precursors (source) into many chemical species leading to plasma formation. This plasma comprises a complex mixture of electrons, ions and neutral radicals. The electrons due to their much higher mobilities respond to the rapidly changing potential. As a result, a negative potential is formed at the substrate surface, which in turn accelerates positive ions towards the substrate resulting in ion sputtering of the substrate to release substrate atoms. The ions are typically accelerated up to relatively low voltages in the range 20–50 V due to the grounded anode being much larger in area compared to the powered cathode. Ions colliding with neutral radicals result in the creation of energetic neutrals. These neutrals are accelerated towards the substrate surface where they

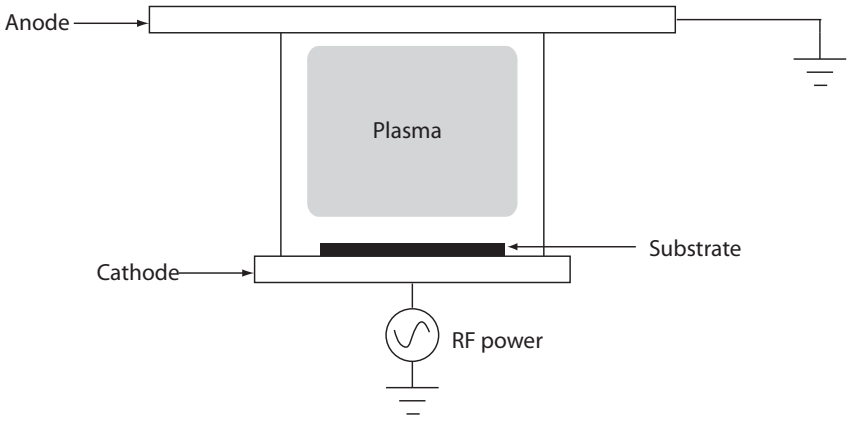


Figure 3.17: Schematic of a reactive ion etching (RIE) dry etch tool.

react with the surface to form highly volatile compounds and in the process chemically etch the surface.

The separate processes can be made to be dominant during etching by tuning the process parameters [146]. At high pressures and low voltages chemical etching dominates, this leads to low anisotropy. The material etch is more selective, and the etch damage is lower due to the decrease in ion bombardment resulting from the reduced voltage. When the parameters are reversed, the physical process dominates. The result is highly anisotropic etch profiles with near vertical side walls, however the material selectivity is reduced and more significantly the surface suffers from increased sputter damage. The low plasma pressure further reduces the etch rate. This can be quickened by increasing the ion and radical concentrations in the plasma as this has a direct effect on the etch rate. Although the etch rate can be increased by supplying higher RF power for increased dissociation, the same RF power is used to accelerate the ions onto the substrate. The corresponding increase in ion energies also leads to more surface damage. Independent control of the plasma dissociation and ion energy is provided for by the ICP technique, similar to that described in Section 3.4.2.1. As a result, fast etch rates and highly anisotropic etch profiles are achieved, while damage to the substrate is minimised.

Dry-etching is well suited for etching features of small geometries with run-to-run reproducibility. The obtained etch profiles are nearly independent of the crystallographic orientation of the substrate. However, there are a couple of disadvantages associated with dry-etching. For one, it is a more complex process compared to wet etching. Moreover, the physical part of the etch process can result in the heating and erosion of the resist layer, which in turn can cause device failure [146].

3.6 Chapter Summary

This chapter outlined the fabrication techniques that shall be used in the III-V MOS device flows, and development of process modules presented in this work. Lithographic techniques were described, with emphasis on EBL due to its exclusivity for MOSFET fabrication in this project. In particular, factors limiting pattern resolution and fidelity in an EBL process were outlined. Thin film deposition and removal techniques, as relevant to III-V MOS device realisation, were discussed. The basics of each technique were explained following which their merits and demerits were highlighted. Attention was drawn to the applicability of these techniques to III-V processing.

4

Characterisation and Metrology

4.1 Introduction

Material design, process module development and integration into a robust fabrication flow are the essential stages of transistor development. Paramount to this are the techniques of characterisation and metrology, which provide the necessary means to experimentally probe and monitor each stage of the development process and the final device. These techniques allow for the various material, process and device metrics to be studied. As a result, invaluable insight into the underlying physical and electrical phenomena can be gained and benchmarks for device performance can be drawn. Pertinent to this are well-established and accurate methods of characterisation and metrology. In this chapter the techniques used to characterise the semiconductor material, the contacts made to it, the gate stack and the completed device are outlined. Specifically, the Hall effect and van der Pauw techniques for carrier density and mobility extraction, and the transmission line model for extracting contact resistance are described. Attention is then drawn to the characterisation of MOS devices, in particular the extraction procedures used to derive key figures of merit.

4.2 Material Characterisation

The performance potential of devices originates from the transport properties at the material level, defined by the carrier density and the mobility of the channel. Since these parameters directly impact the drain current behaviour of transistors, the eventual device performance largely depends on the quality of the starting epi-layers. Layer design optimisation, complemented by high-quality growth, are essential for obtaining the desired

quality of the material. Important to this process is the technique of Hall effect measurement which provides the means to determine channel carrier concentration and mobility. Due to the simplicity of implementation and ease of usage, this is a widely used method for semiconductor material characterisation.

The Hall effect principle describes the behaviour of charge carriers in a semiconductor under the influence of both electric and magnetic fields [147]. Considering a p-type semiconductor, holes drifting with a velocity (v_x) under the force of an applied longitudinal electric field (E_x) will experience a Lorentz force (F_L) in the presence of a transverse magnetic field (B_z) as shown in Figure 4.1. This force, directed perpendicular to both the magnetic field and the motion of the charge carriers, is given as:

$$F_L = -q\, v_x\, B_z\,. \tag{4.1}$$

From the resulting deflection, holes accumulate on one side of the semiconductor slab creating a net positive charge, whilst a net negative charge appears on the opposite side due to hole depletion. This gives rise to a transverse electric field (E_y), also known as the Hall field, which directly opposes the Lorentz force. When the force associated with the Hall field balances the Lorentz force the deflection of holes ceases. The Hall field required to attain this state of equilibrium in turn establishes the Hall voltage:

$$V_H = \frac{R_H\, B_z\, I_x}{t} = \frac{B_z\, I_x}{tqp}\,, \tag{4.2}$$

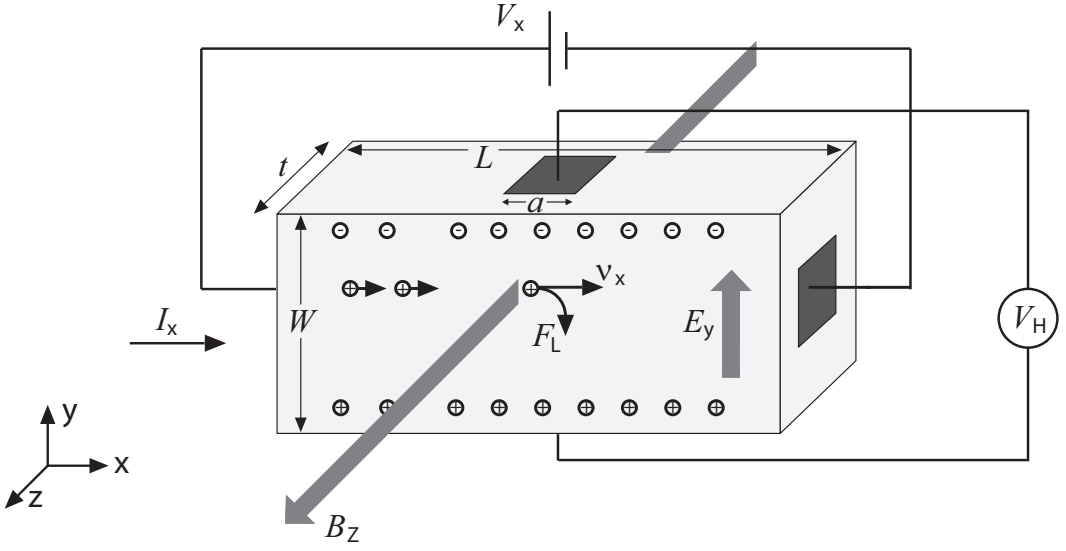


Figure 4.1: Schematic depicting the Hall effect in a p-type bar-shaped semiconductor sample.

where I_x is the hole drift current, t is the semiconductor thickness, p is the channel carrier concentration, q is the electronic charge and R_H=$1/qp$ is the Hall coefficient. Therefore for a semiconductor of known dimensions and for a given magnetic field, measurements of the drift current and the Hall voltage would enable the hole density to be calculated from Equation (4.2).

Knowledge of the Hall coefficient also allows for the determination of the hole drift mobility (μ_H), also know as Hall mobility, according to:

$$\mu_H = \frac{R_H}{\rho}, \tag{4.3}$$

where ρ is the electrical resistivity at zero magnetic field. The resistivity can be evaluated from measurements of the applied potential (V_x), the resulting drift current and semiconductor dimensions as follows:

$$\rho = \frac{V_x W t}{I_x L}, \tag{4.4}$$

where L and W represent the length and width of the semiconductor slab.

A Hall bar structure, akin to that shown in Figure 4.1, is a common sample geometry for Hall effect measurements. In the design of Hall bars, the geometrical layout needs to be carefully considered as this directly affects the measurement accuracy. One geometrical consideration concerns the tendency of the end contacts to short out the Hall voltage and is related to the ratio of the sample length to width (L/W). For a 4-contact Hall bar (Figure 4.1), the deviation of the measured to the actual Hall voltage is less than 1% if L/W=3 whereas the deviation increases to 30% if L/W=1 [148]. Another geometrical consideration relates to the contacts for Hall voltage measurements. Although ideally the contacts should be point-like with vanishing length, in reality they are of a finite size. This has implications for the Hall mobility and the associated degradation can be approximated, for $\mu_H B_z \ll 1$, by [149]:

$$\frac{\Delta\mu_H}{\mu_H} = 1 - \left(1 - e^{-\pi L/2W}\right)\left(1 - \frac{2a}{\pi W}\right), \tag{4.5}$$

where $\Delta\mu_H$ is the increase required in the Hall mobility to obtain the true value and a is the contact length. Assuming L/W= 3, in the case of a point contact, the degradation in mobility is negligible (0.9%). However for finite contacts, assuming a/W=0.2, the mobility degradation becomes appreciable (13%). The addition of sidearms, as illustrated in Figure 4.2 whereby the contacts are located at the ends of the arms, can circumvent

the contact size issue [150]. It is also noted that the sample geometry has been modified from a 4-contact to a 6-contact Hall bar. The benefit of this configuration is that the parallel voltage (V_x) can now be measured between two sidearm contacts. The change in geometry also necessitates the modification of the aspect ratios. The sample aspect ratio requirements for a 6-contact Hall bar [149] are indicated in Figure 4.2. By keeping to these ratios, the fabricated Hall bar would compare closely to the ideal case of a rectangular sample with true point contacts, thereby ensuring accuracy of measurements.

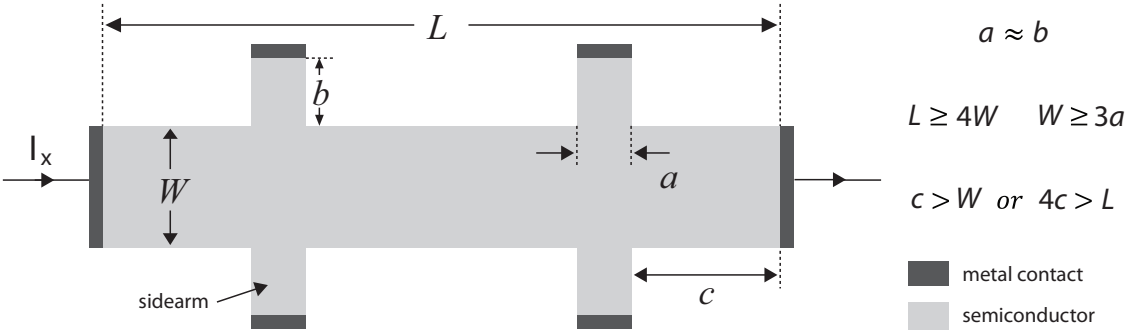


Figure 4.2: A top view schematic of a 6-contact Hall bar with sidearms. Aspect ratios required of the sample geometry are also shown.

The accuracy of Hall effect measurements and subsequent parameter extraction, as noted above, is sensitive to sample geometries that need to be precisely known. An alternative technique, proposed by van der Pauw [151], obviates this geometrical dependence such that the Hall coefficient and resistivity can be evaluated from a sample of arbitrary shape with contacts located arbitrarily around its circumference, as shown in Figure 4.3(a). Then, the resistivity of the sample, ρ, with B_z=0, is given by:

$$\rho = \frac{\pi t}{\ln 2} \frac{(R_{\mathrm{AB,CD}} + R_{\mathrm{BC,DA}})}{2} F = \frac{\pi t}{\ln 2} \frac{1}{2} \left(\frac{V_{\mathrm{CD}}}{I_{\mathrm{AB}}} + \frac{V_{\mathrm{DA}}}{I_{\mathrm{BC}}} \right) F \,. \tag{4.6}$$

$R_{\mathrm{AB,CD}}$ and $R_{\mathrm{BC,DA}}$ are the resistances obtained from measuring the potential drop across two adjacent contacts due to current injection into the other adjacent contact pair. F is a correction factor that is associated with the difference between $R_{\mathrm{AB,CD}}$ and $R_{\mathrm{BC,DA}}$, and can be calculated from a transcendental equation [151]. For a sample of symmetric geometry and contact placement, as shown in Figure 4.3(b), $R_{\mathrm{AB,CD}}$=$R_{\mathrm{BC,DA}}$=R resulting in a correction factor of 1. In this case, Equation (4.6) can be simplified as:

$$\rho = \frac{\pi t R}{\ln 2} \,. \tag{4.7}$$

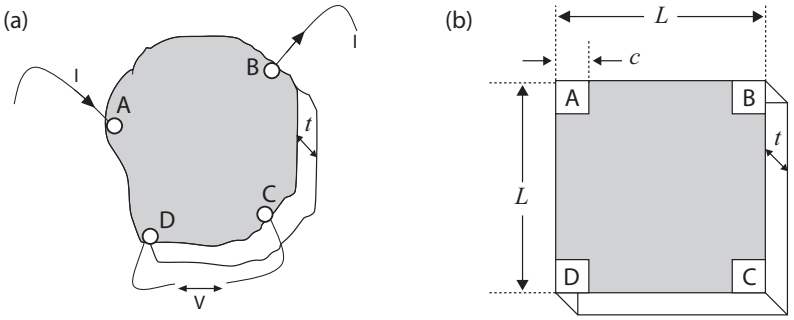


Figure 4.3: Van der Pauw structures: (a) a sample of arbitrary geometry with contacts arbitrarily located around the periphery and (b) a sample of symmetric geometry with contacts located symmetrically around the periphery.

By eliminating the thickness dependence from Equation (4.7) the sheet resistance (R_{sh}), a useful measure of conductivity relating to two-dimensional carrier transport, can be obtained as follows:

$$R_{\mathrm{sh}} = \frac{\rho}{t} = \frac{\pi R}{\ln 2}\,. \tag{4.8}$$

The determination of the Hall coefficient is based on measuring the change in the resistance (ΔR) associated with a diagonal contact pair, from changing the magnetic field, with one measurement usually taken at zero field and the other at some finite field value:

$$\Delta R = R_{\mathrm{BD,AC}}\,(B_{\mathrm{z}} \neq 0) - R_{\mathrm{BD,AC}}\,(B_{\mathrm{z}} = 0)\,. \tag{4.9}$$

$R_{\mathrm{BD,AC}}$ is measured by injecting current across contacts A and C, diagonally located, with the resulting potential drop measured across the other diagonal pair of contacts B and D. From the change in resistance due to a change in the magnetic field, the Hall coefficient can be calculated as:

$$R_{\mathrm{H}} = \frac{t}{B_{\mathrm{z}}}\,\Delta R\,. \tag{4.10}$$

The Hall mobility, calculated from the product of the conductivity and Hall coefficient (see Equation (4.3)), is expressed as:

$$\mu_{\mathrm{H}} = \frac{\ln 2}{\pi B_{\mathrm{z}}}\,\frac{\Delta R}{R}\,. \tag{4.11}$$

It is worth emphasising that similar to the sheet resistance evaluation, sample thickness is not required in determining the Hall mobility. Using the relation $R_{\mathrm{sh}} = 1/p_{\mathrm{s}} q \mu_{\mathrm{H}}$, the sheet carrier density can then be solved as:

$$p_{\mathrm{s}} = \frac{B_{\mathrm{z}}}{\Delta R\, q}\,. \tag{4.12}$$

The accuracy of the measurements obtained with the van der Pauw technique is also affected by geometrical errors. Although point-like contacts are assumed in the analysis, this is impossible to achieve in practice. Thus, similar to the Hall bar structure, the contacts are of a finite size in the fabricated van der Pauw structure. For samples and contacts based on a square geometry (Figure 4.3(b)), the corrections for finite size contacts is determined from the aspect ratio of the contact length to the sample length (c/L). The resistivity correction factor $\Delta\rho/\rho$ is approximately proportional to $(c/L)^2$ whilst the Hall coefficient correction factor $\Delta R_{\mathrm{H}}/R_{\mathrm{H}}$ is proportional to c/L [149]. For an aspect ratio of 1/10, the correction for the resistivity only amounts to 1%. The correction for the Hall coefficient however is more severe ($\Delta R_{\mathrm{H}}/R_{\mathrm{H}}$=10%). This means the extracted Hall mobility and sheet carrier density will be underestimated and overestimated respectively by the same percentage margin. This can be mitigated by employing samples with a larger geometry and/or reducing the contact size. However the constraints associated with measuring using wafer probes set the lower limit of the contact size, which was $150\,\mu$m$\times150\,\mu$m. In this work, both the Hall bar and van der Pauw structures were used.

4.3 Contact Characterisation

In addition to the quality of the epitaxially grown active layers, another fundamental performance limiting factor of transistors is the quality of the contacts made to the semiconductor material. As outlined in Section 2.5, high contact resistance due to the poor quality of ohmic contacts detrimentally affects transistor performance by limiting access to the materials intrinsic capability. The ability to characterise ohmic contacts allows for their performance to be evaluated and optimised. This is key to maximising access to the potential of the material for enhanced device performance.

The transmission line model (TLM), also known as the transfer length method, is a technique that is widely used to characterise the performance of metal contacts to planar semiconductor devices. The approach, originally proposed by Shockley [152], is based on a test structure comprising a linear array of identical contacts of width W and length L, spaced by progressively increasing distance d as shown in Figure 4.4(a). The total resistance (R_{T}) between two adjacent contacts is determined by forcing a current between the contacts using a pair of probes and measuring the resulting potential drop by a second pair of probes. This should result in the measured resistance being proportional to the spacing between the contacts (Figure 4.4(b)), the linear relationship is expressed

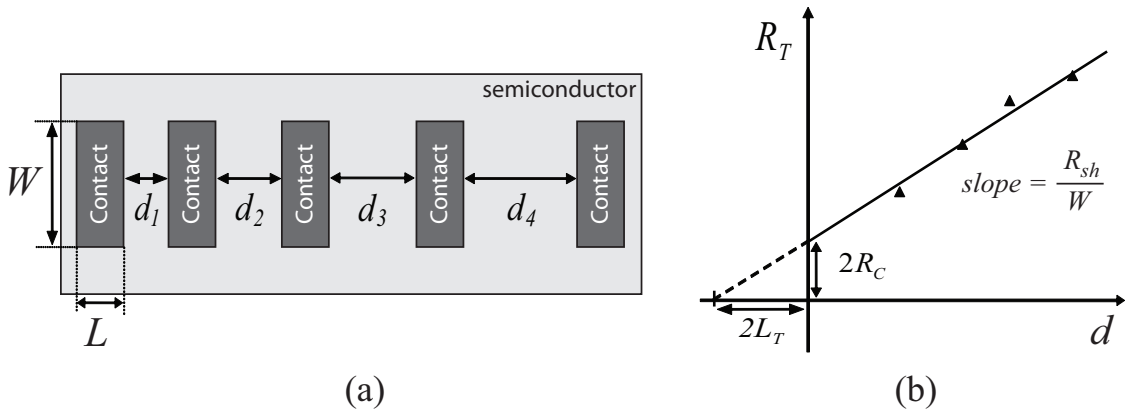


Figure 4.4: (a) A schematic diagram of the test structure for TLM measurements and (b) the resulting plot of total resistance as a function of contact spacing for parameter extraction.

as:

$$R_{\mathrm{T}} = \frac{R_{\mathrm{sh}}}{W}\, d + 2\, R_{\mathrm{c}}\,. \tag{4.13}$$

The slope of the linear regression gives the semiconductor sheet resistance (R_{sh}), a normalised quantity expressed in Ω/sq. Extrapolating the regression to zero contact spacing results yielding the resistance of two planar contacts. The contact resistance (R_{c}), usually normalised to a contact width of 1 mm, represents the resistance of one contact with units of Ω.mm.

Further extrapolation of the regression to $R_{\mathrm{T}} = 0$ yields twice the value of the transfer length (L_{T}), distance from the contact edge over which $\frac{1}{e}$ of the total current flows from the semiconductor to the contact [153]. For efficient ohmic behaviour, the contact length needs to be at least twice as long as the transfer length ($L \geq 2L_{\mathrm{T}}$), otherwise there is a significant increase in contact resistance due to current crowding effects [154].

Another figure of merit quite often used as a measure of the quality of contacts is the specific contact resistivity (ρ_{c}) defined as:

$$\rho_{\mathrm{c}} = \left(\frac{\delta J}{\delta V}\right)^{-1}_{V=0}, \tag{4.14}$$

where V is the potential drop across the contact and J is the current density through the contact. The current density depends on the current transport mechanisms discussed in Section 2.5.2. The specific contact resistivity can be evaluated from the extracted transfer length in conjunction with the semiconductor sheet resistance ($\rho_{\mathrm{c}} = R_{\mathrm{sh}}\, L_{\mathrm{T}}^2$) [155].

The extraction of L_T and hence ρ_c are based on the assumption that the sheet resistance of the material outwith the contacts is identical to that directly below the contacts. While this is a valid assumption for as-deposited metal contacts, the same does not apply to sintered contacts due to alloying effects [155]. In this case, not directly accounting for the sheet resistance underneath the contacts (R_{sk}) would result in inaccurate values of L_T and ρ_c. The refinements to the TLM method taking into account the alloying effects are reported in [155]. This refined model for evaluating ρ_c was not used since the task of ascertaining R_{sk} is not trivial. To avoid inaccuracies contact performance in this work is reported in terms of the contact resistance in Ω.mm.

4.4 MOS Device Characterisation

4.4.1 MOS Capacitor Characterisation

C–V measurements are a staple in the electrical characterisation of MOS devices. These measurements, performed on MOS capacitors, provide invaluable insight into the quality of the gate dielectric and the interface its forms with the semiconductor, which has a direct impact on transistor performance. Non-ideal C–V behaviour can be directly observed and quantified from the measured characteristics. It is therefore a vital tool in understanding the deviations of the fabricated MOS capacitor from its ideal characteristics. Important defect and physical information relating to the dielectric/semicondutor stack can be evaluated, enabling the process and material parameters that determine device operation to be interrogated. This section describes the techniques for extracting the flatband voltage, metrics relating to the non-ideal C–V behaviour and interface trap density.

4.4.1.1 Flatband Voltage

The flatband voltage (V_{fb}), as previously described in Section 2.3.1, refers to the applied bias for which there is zero electric field across the gate dielectric, and thus no band bending at the dielectric/semiconductor interface (see Figure 2.7). Flatband voltage is used extensively to analyse MOS operation. For instance the V_{fb} shifts of the C–V curves according to the different metal gate work functions are often used as one indicator of an unpinned Fermi level [29, 39, 41]. Moreover, physical parameters relating to MOS devices (e.g. oxide charges) can be extracted using V_{fb} [156, 157].

Flatband voltage can be determined by a number of methods, all of which rely on a C–V characteristic measurement of a MOS capacitor. A high frequency measurement is

typically used to minimise the capacitance contribution of D_{it} to the C–V characteristic [157]. The first method involves a comparison between the simulated C–V characteristic of an ideal MOS capacitor with the experimental C–V [156]. In the ideal case, the flatband condition is achieved under zero applied bias (V_{fb}=V_{g}=0 V) which coincides with the flatband capacitance (C_{fb}). Thus, V_{fb} is a measure of the shift between the ideal and experimental curves at C_{fb}, which is schematically illustrated in Figure 4.5(a). In the second method [156], V_{fb} is determined from the value of C_{fb} calculated using the relation:

$$C_{\mathrm{fb}} = \frac{C_{\mathrm{ox}}\, C_{\mathrm{sfb}}}{C_{\mathrm{ox}} + C_{\mathrm{sfb}}}, \tag{4.15}$$

where C_{sfb} is the semiconductor capacitance under flatband conditions and C_{ox} is the gate dielectric capacitance (see Equation (2.30)). The C_{sfb} value is calculated as follows:

$$C_{\mathrm{sfb}} = \frac{\epsilon_{\mathrm{s}}\, \epsilon}{\lambda_{\mathrm{d}}}, \tag{4.16}$$

where ϵ_{s} is the relative permittivity of the semiconductor, ϵ is the permittivity of free space and λ_{d} is the Debye length (see Equation (2.25)). V_{fb} is then graphically determined as the gate voltage corresponding to the computed C_{fb} from the measured C–V characteristic as shown in Figure 4.5(b).

The aforementioned methods necessitate precise knowledge of the material and design parameters of the fabricated MOS structure. For instance C_{max} in accumulation, in theory, equates to C_{ox} and therefore can be used for determining the flatband capacitance. However, in real MOS systems C_{max} in accumulation is reduced from the dielectric capacitance due to the effects of DOS and charge quantisation in semiconductors [99] (see Section 2.3.4.1), which would result in the incorrect evaluation of C_{fb}. In this case additional metrology would be required to determine C_{ox}, for example transmission electron microscopy (TEM). This would provide an accurate measurement of the physical dielectric thickness. In addition, the dielectric constant can be accurately ascertained from the slope of the experimental CET versus physical dielectric thickness as determined from TEM [157]. Flatband voltage extraction via a purely experimental technique would therefore be advantageous. Hillard et al. [158] present such a method in which the lower knee of the $(1/C_{\mathrm{HF}})^2$ versus V_{g} plot, derived from the experimental C–V measurement, corresponds to V_{fb} as illustrated in Figure 4.5(c). However, determining the exact position of the knee can be difficult. A first order differential of the $(1/C_{\mathrm{HF}})^2$ curve would produce a second curve (see Figure 4.5(c)), for which the maximum slope of the left flank corresponds to V_{fb}. To accurately determine the maximum of this slope, a second

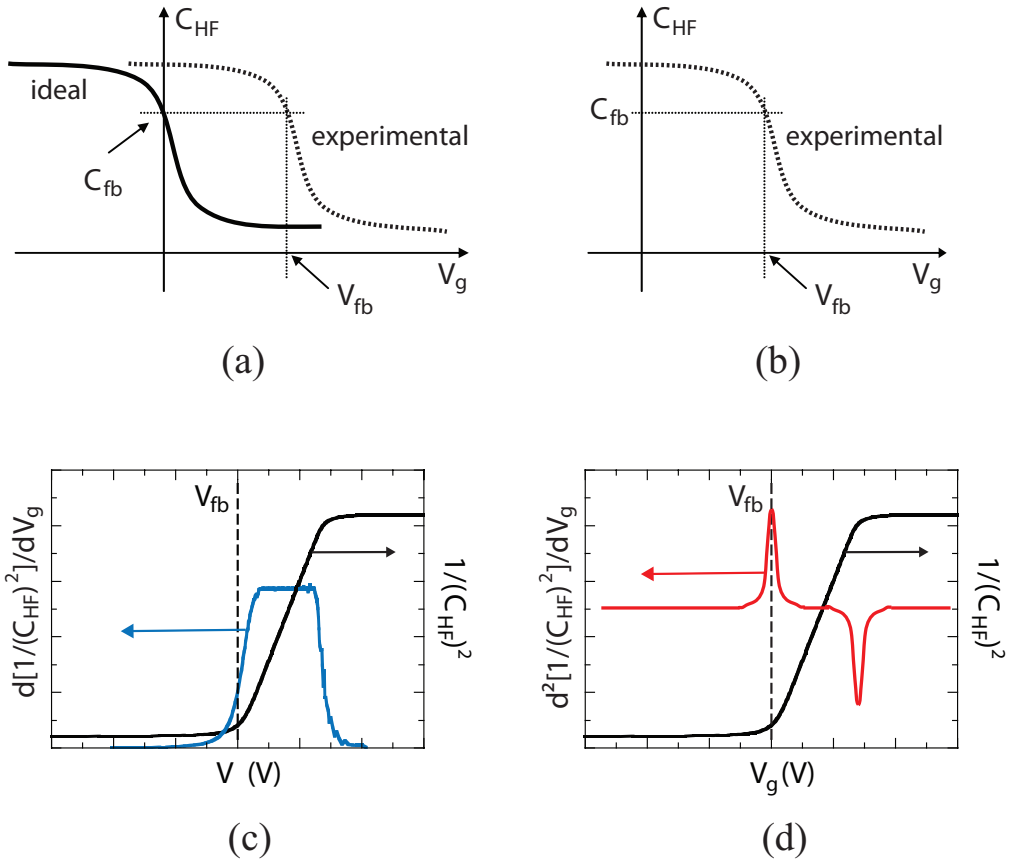


Figure 4.5: Methods for flatband voltage (V_{fb}) determination of a n-MOS structure: (a) Shift between the ideal and experimental C–V curves at flatband capacitance (C_{fb}), determined from the ideal curve at $V_{\mathrm{g}} = 0$, yields V_{fb}, (b) the flatband capacitance, calculated based on the physical parameters of the MOS structure, is used to extract V_{fb} from the experimental C–V characteristic, and the Hillard method [158] in which the flatband voltage coincides with (c) the maximum slope of the left flank of the first order differential and correspondingly (d) the maximum peak of the second order differential of the $1/(C_{\mathrm{HF}})^2$ versus V_{g} plot.

differentiation is usually performed. This yields a sharply peaked curve, the maximum of which corresponds to V_{fb} (see Figure 4.5(d)). This method was adopted for flatband voltage determination in this work.

4.4.1.2 Non-Ideal C-V Behaviour

Charge defects in the gate stack tend to exact a detrimental influence on the transistor metrics of drive current, threshold voltage, subthreshold slope, mobility and transconductance, resulting in a degradation of device performance. The effect of these defects

is also directly observed from the measured C–V responses of a MOS structure as deviations from its ideal characteristics; the physics of which was previously described (see Section 2.3.4.3). A pictorial representation of a non-ideal C–V behaviour of a n-MOS capacitor is shown in Figure 4.6, with the metrics relating to the deviations identified. The reduction of the maximum capacitance value with increasing frequency gives form to frequency dispersion in the accumulation region. This metric is defined as:

$$\text{Frequency Dispersion } (\%/\text{dec}) = \frac{C_{\text{LF,max}} - C_{\text{HF,max}}}{C_{\text{LF,max}}} \times \frac{100\%}{N_{\text{dec}}}, \tag{4.17}$$

where $C_{\text{LF,max}}$ and $C_{\text{HF,max}}$ represent the maximum accumulation capacitance of the low frequency and high frequency C–V data, and N_{dec} is the number of frequency decades between the two frequency measurements. The metric of frequency dispersion in depletion is defined as the frequency dependent flatband voltage shift and is given by the relation:

$$V_{\text{fb}} \text{ shift } (\text{mV}) = (V_{\text{LF,fb}} - V_{\text{HF,fb}}) \,, \tag{4.18}$$

where $V_{\text{HF,fb}}$ and $V_{\text{LF,fb}}$ represent the flatband voltages of the high and low frequency curves. The procedure for calculating this metric is as follows: the method of Hillard et al. [158], presented in the previous section, is used in extracting $V_{\text{hf,fb}}$. Then C_{fb} is determined from the high frequency C–V curve using the extracted $V_{\text{HF,fb}}$ value. Following this, the gate bias corresponding to the previously determined C_{fb} is extracted from the low frequency curve and this value is taken as $V_{\text{LF,fb}}$.

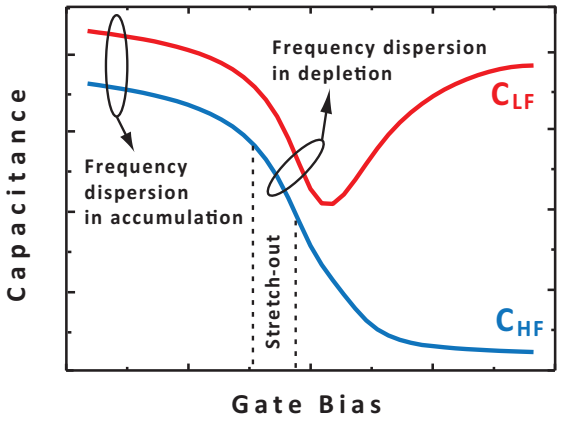


Figure 4.6: Non-ideal C–V behaviour of a n-MOS capacitor resulting from the effects of interface defects.

The final metric of stretch-out is defined as:

$$\text{Stretch-out } (\text{F/cm}^2\text{V}) = \frac{\Delta C_{\text{HF}}}{\Delta V_{\text{g}}}. \tag{4.19}$$

The slope of the C–V curve is extracted between V_{fb} and V_{fb}+0.3 V.

Lower frequency dispersion in accumulation, smaller V_{fb} shifts and reduced stretch-out (larger C–V slope) all point to lower interface defects. While these metrics provide an indication on the quality of the dielectric/semiconductor interface, it cannot be used directly to obtain D_{it}. Regardless, the metrics serve as an invaluable qualitative assessment tool when drawing comparisons between C–V characteristics of MOS capacitors subject to different process conditions, ranging from dielectric deposition techniques to interface control technology.

4.4.1.3 Interface Trap Density

The techniques most commonly used to determine interface trap density, D_{it}, in MOS capacitors are based on frequency dependent conductance or capacitance measurements.

Conductance Measurements

One of the most sensitive methods for D_{it} extraction is the conductance method [159], based on measuring the equivalent parallel conductance (G_{p}) as a function of frequency for different gate biases. Conductance peaks are representative of energy losses associated with changes in the interface trap occupancy and therefore provide a direct measure of D_{it}. The ability to probe D_{it} levels in the order of $10^9\,\text{cm}^{-2}\text{eV}^{-1}$ in depletion and weak inversion, and derive information on the majority carrier capture cross section and surface potential fluctuations, makes this a popular method [156]. However the conductance method, originally developed for SiO_2/Si interfaces, has several limitations that need to be considered in the context of high-k/III-V interfaces. In the standard conductance method, the capture cross section and D_{it} are assumed to be independent of the trap energy, resulting in symmetrical G_{p}–ω peaks. While this serves as a good approximation for the SiO_2/Si interface, the conductance peaks tend to exhibit a strong asymmetry for interfaces between high-k dielectrics and III-V semiconductors. For the conductance method to be applicable to high-k/III-V interfaces, the non-uniform distribution needs to be accounted for [160]. Another limitation is the ineffectiveness of the method for interfaces with large D_{it}, which is typical of high-k/III-V interfaces. In particular for $qD_{\text{it}} > C_{\text{ox}}$ the oxide capacitance dominates the measured conductance resulting in the underestimation of D_{it}. With increasing qD_{it} the underestimation becomes more severe

and eventually for $qD_{\mathrm{it}} > 4C_{\mathrm{ox}}$ the parallel conductance is no longer sensitive to D_{it} [161–163]. A final source of error concerns conductance in weak inversion, which increases due to minority carrier generation-recombination. Not accounting for this, and solely attributing the conductance to interface trap response, results in the overestimation of D_{it} [160].

Capacitance Measurements

The most common capacitance based techniques for D_{it} extraction are the high frequency (Terman) [164], low frequency (Berglund) [165] and high-low frequency (Castagné-Vapaille) [166] methods. In the Terman method, a high frequency C–V measurement is compared against an ideal C–V curve modelled at the same frequency. At sufficiently high frequencies, the interface traps do not capacitively contribute to the measurement ($C_{\mathrm{it}} = 0$). The only effect of interface traps arises from the change in trap occupancy with the slowly varying DC bias, which in turn induces a stretch-out of the C–V along the gate bias axis. This difference between the measured and modelled curves is used to evaluate D_{it}. In contrast to the Terman method, D_{it} extraction by means of the Berglund method is based on a low frequency C–V measurement. For a sufficiently low frequency all traps are able to follow the ac signal, and therefore the measured C–V profile contains an additional C_{it} contribution. Similar to the Terman method, trap density is evaluated from a comparison between the measured and modelled C–V profiles.

The main advantage of the Terman and Berglund methods is that trap densities can be probed over a large portion of the bandgap [167]. A considerable drawback of both methods however is the need for an ideal C–V curve which when not modelled accurately, due to inaccuracies in process and material parameters (e.g. doping density), will result in erroneous D_{it} [160]. Castagné and Vapaille proposed a rather simplified alternative of combining the high and low frequency C–V measurements, which precludes theoretical models and its associated inaccuracies, in extracting the trap density. By comparing the low frequency response ($C_{\mathrm{it}} \neq 0$) to the high frequency response ($C_{\mathrm{it}} = 0$) and using the relation $C_{\mathrm{it}} = qD_{\mathrm{it}}$, the interface trap density can be expressed as [166]:

$$D_{\mathrm{it}}(V_{\mathrm{g}}) = \frac{C_{\mathrm{ox}}}{q}\left(\frac{C_{\mathrm{LF}}/C_{\mathrm{ox}}}{1 - C_{\mathrm{LF}}/C_{\mathrm{ox}}} - \frac{C_{\mathrm{HF}}/C_{\mathrm{ox}}}{1 - C_{\mathrm{HF}}/C_{\mathrm{ox}}}\right). \tag{4.20}$$

Unlike the Terman and Berglund methods, the high-low method is only valid over a limited portion of the bandgap, which extends from the onset of inversion to a surface potential towards the majority carrier band edge [166]. In n-MOS capacitors, only trap levels in the lower half of the bandgap can be probed. By complementing this with trap

levels extracted from a p-MOS capacitor, D_{it} over a large portion of the bandgap can be extracted.

The accuracy of the capacitance based methods hinges on being able to obtain "true" high and low frequency C–V measurements [160]. If the C_{it} contribution to the low frequency profile is not representative of all interface traps then D_{it} is underestimated. This can be minimised by measuring to as low a frequency as possible. For interfaces with a large D_{it} high frequency measurements are not necessarily devoid of AC contributions from interface traps, even at a frequency of 1 MHz [156], which would then underestimate D_{it}. One way of minimising the AC response to the high frequency curve is to perform the measurement at reduced temperatures (<300 K) [26]. Additionally D_{it} extraction based on all the aforementioned methods is sensitive to C_{ox}. An overestimation of C_{ox} would result in the underestimation of D_{it}. As previously highlighted, the effects of DOS and charge quantisation renders C_{ox} extraction from the maximum accumulation capacitance erroneous [99, 160]. An accurate C_{ox} value should therefore be obtained from an independent measurement.

In this work, the high-low frequency method was adopted for D_{it} extraction as it avoids many of the problems associated with the Terman, Berglund and conductance methods. Since C–V profiles are measured as a function of gate bias, D_{it} is also extracted as a function of bias using the high-low method. It is more useful to express D_{it} as a function of energy position within the semiconductor bandgap. In order to do this, a relationship between the surface potential (ψ_{s}) and gate bias is needed. Using the Berglund integral [165] the ψ_{s}–V_{g} relation can be calculated from the low frequency capacitance measurement (C_{LF}) as follows:

$$\psi_{\mathrm{s}}(V_{\mathrm{g2}}) - \psi_{\mathrm{s}}(V_{\mathrm{g1}}) = \int_{V_{\mathrm{g1}}}^{V_{\mathrm{g2}}} \left(1 - \frac{C_{\mathrm{LF}}(V_{\mathrm{g}})}{C_{\mathrm{ox}}}\right) dV_{\mathrm{g}}\,, \tag{4.21}$$

where V_{g1} can be any voltage for which the surface potential is accurately known. Since there is no band bending at the flatband condition, setting $V_{\mathrm{g1}} = V_{\mathrm{fb}}$ would make $\psi_{\mathrm{s}}(V_{\mathrm{g1}}) = 0$ in Equation (4.21). The integral should then be split into two parts; one integrating from flatband to accumulation and another from flatband to weak inversion. From the obtained ψ_{s}–V_{g} relation, the trap energy position (E_{tr}) at the dielectric/semiconductor interface can be calculated as follows:

$$E_{\mathrm{tr}} = \frac{E_{\mathrm{g}}}{2} + q\psi_{\mathrm{s}} - q\psi_{\mathrm{B}}\,, \tag{4.22}$$

where ψ_{b} is the Fermi level position with respect to the intrinsic Fermi level in the bulk.

4.4.2 MOSFET Characterisation

Current-voltage (I–V) measurements are routinely obtained from MOSFETs to study their electrical properties. The measured output and transfer characteristics provide useful information on device performance which is intrinsically linked to the quality of the semiconductor material, source-drain contacts, gate dielectric, dielectric/semiconductor interface and device processing. The I–V profiles provide a means of extracting and comparing the d.c. metrics of the device. While the extraction of certain metrics are straightforward (e.g. saturation drain current), others may require additional analysis (e.g. effective mobility). Some of these metrics can also be used to derive additional information. For instance the quality of the dielectric/semiconductor interface can be assessed directly from the subthreshold slope. In this section, the experimental techniques used in extracting the threshold voltage, source/drain series resistance, effective mobility and interface trap density are outlined.

4.4.2.1 Threshold Voltage

V_{th} is a fundamental parameter of importance that signifies the turn-on point of transistors. Knowledge of V_{th} is crucial for reliably comparing the performance of devices. Moreover, the relevant V_{th} tuning measures, for example the choice of gate metal, can be employed to achieve the desired transistor operation; depletion-mode (D-mode) or enhancement-mode (E-mode), the latter being a pre-requisite for digital logic. In conventional MOSFETs the gate voltage at which the channel crosses over from being weakly to strongly inverted (minority carriers determine the drain current) is defined to be V_{th}. This voltage corresponds to the condition $\psi_{\mathrm{s}} = 2\psi_{\mathrm{b}}$. However this may not be an adequate description for modern devices, including the flatband MOSFET, which operates on the basis of majority carriers [94]. A more practical concept of threshold can be defined as V_{g} required to turn the channel from being weakly to strongly conductive between the source and drain.

While several extraction techniques have been proposed to determine V_{th} [168], the two most commonly adopted for analysis and benchmark of device performance are the constant-current (CC) and the linear extrapolation (LE) methods. CC is the method that is widely used in industry and is an ideal choice for well-established technologies. In this empirical method, V_{th} is defined as V_{g} corresponding to an assigned arbitrary constant drain current and $V_{\mathrm{d}} \leq 100\,\mathrm{mV}$. This value is typically evaluated as $I_{\mathrm{d}} = (W_{\mathrm{g}}/L_{\mathrm{g}}) \times 10^{-7}$ (see Figure 4.7(a)). Although preferred for its simplicity the main drawback of this method

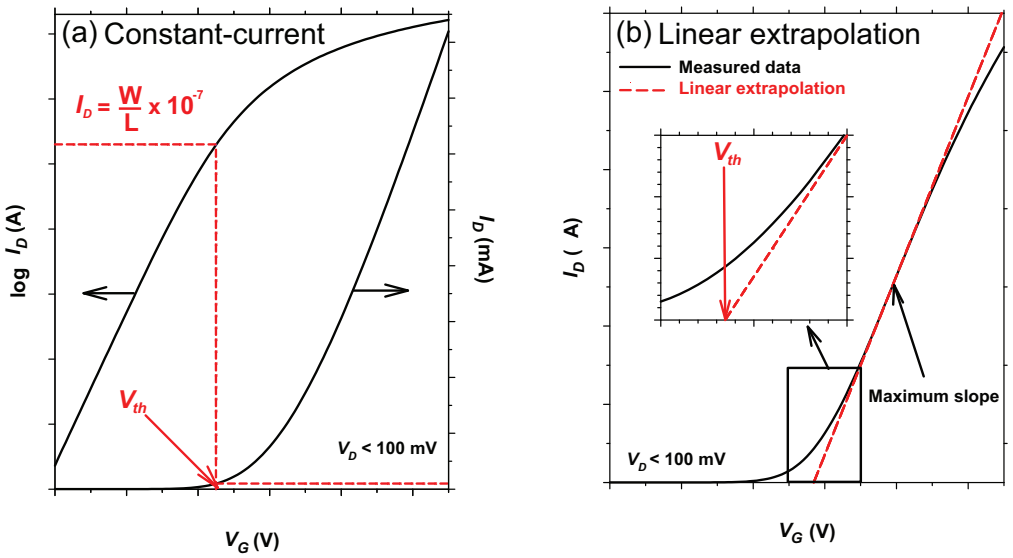


Figure 4.7: Threshold voltage extraction in the linear region using (a) constant-current and (b) linear extrapolation. Adapted from [168].

is the arbitrary nature with which I_d is selected, which in turn affects the outcome of V_th leading to inconsistent results [168].

LE is arguably the most widely adopted experimental technique for V_th extraction. In the linear region ($V_\mathrm{d} \leq 100\,\mathrm{mV}$), there is a linear dependence between I_d and V_g (see Equation (2.4)). Therefore, a linear extrapolation of the I_d–V_g plot from its maximum slope to $I_\mathrm{d} = 0$, enables the V_g-axis intercept to be obtained as shown in Figure 4.7(b). From this, V_th can be determined by accounting for the $V_\mathrm{d}/2$ term in Equation (2.4). The slope of the plot ($dI_\mathrm{d}/\mathrm{d}V_\mathrm{g}$) corresponds to the transconductance and its maximum coincides with point at which g_m is maximum. In the saturation region ($V_\mathrm{d} \geq V_\mathrm{g} - V_\mathrm{th}$), there is a square-law dependence between I_d and V_d (see Equation (2.6)). As with the linear region, a linear extrapolation of $\sqrt{I_\mathrm{d}}$–V_g at its maximum 1st derivative enables V_th to be determined. For all measured devices of this work, the method based on LE was employed to extract V_th.

4.4.2.2 Source/Drain Series Resistance

The effect of the source-drain series resistance (R_sd) is to reduce the intrinsic bias and degrade the drive current capability of MOSFETs, as highlighted in Section 2.5. By extracting R_sd, its impact on device performance can thus be assessed. In addition it provides a means to decouple the relative contributions of the resistive components to the total series resistance.

The most common technique to extract R_{sd} is to treat the MOSFET as an equivalent circuit comprising three resistances in series (see Figure 2.21): source (R_{s}), drain (R_{d}) and channel (R_{ch}). In principle, R_{sd} ($= R_{\mathrm{s}} + R_{\mathrm{d}}$) is assumed to be a constant and for a given gate bias the channel resistance is proportional to the device gate length (L_{g}). This then allows R_{sd} to be evaluated from an array of identical MOSFETs with different gate lengths [169, 170], similar to the transmission line method. In order to illustrate this extraction technique, the on-resistance ($R_{\mathrm{on}} = V_{\mathrm{d}}/I_{\mathrm{d}}$) as a function of gate length, for arbitrary values, of otherwise identical devices is plotted in Figure 4.8. The on-resistance is evaluated in the linear region (low V_{d}) of device operation and for a high gate bias. Linear regression is performed through the series of data points and interpolated at zero gate length to yield the value of resistance corresponding to R_{sd} as shown in Figure 4.8.

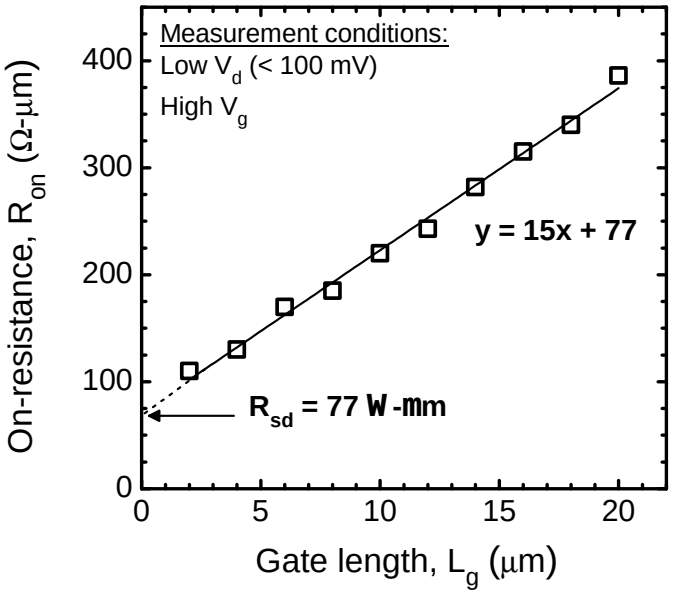


Figure 4.8: Example plot of on-resistance (R_{on}) as a function of gate length (L_{g}), based on arbitrary values, of otherwise identical MOSFETs at a low drain bias (V_{d}) and high gate bias (V_{g}).

4.4.2.3 Channel Mobility

A key metric that defines the performance of MOSFETs is the effective mobility (μ_{eff}) of charge carriers in the channel. The effective mobility can be calculated by considering the drain current expression of a p-channel MOSFET given as:

$$I_{\mathrm{d}} = I_{\mathrm{drift}} + I_{\mathrm{diffusion}} = \frac{W_{\mathrm{g}}\,\mu_{\mathrm{eff}}\,Q_{\mathrm{ch}}\,V_d}{L_{\mathrm{g}}} - W_{\mathrm{g}}\,\mu_{\mathrm{eff}}\,\frac{kT}{q}\,\frac{\mathrm{d}Q_{\mathrm{ch}}}{\mathrm{d}x}\,, \tag{4.23}$$

where Q_{ch} is the hole density in the channel. It is observed that Equation (4.23) is comprised of a drift current (I_{drift}) and a diffusion current ($I_{\mathrm{diffusion}}$). The diffusive term can be neglected if the drain current is measured in the linear region of device operation. At a low drain bias the channel charge that flows between source and drain becomes more uniform resulting in $\mathrm{d}Q_{\mathrm{ch}}/\mathrm{d}x$ going to zero. Thus Equation (4.23) can be simplified and solved for the effective mobility as:

$$\mu_{\mathrm{eff}} = \frac{I_{\mathrm{d}}\, L_{\mathrm{g}}}{W_{\mathrm{g}}\, Q_{\mathrm{ch}}\, V_{\mathrm{d}}}\,. \tag{4.24}$$

It is important that the I_{d}–V_{g} measurement is obtained from a long channel transistor. This is to exclude SCE on the measured characteristics for accurate mobility extraction.

The calculation of μ_{eff} also relies on the determination of Q_{ch} (see Equation (4.24)), for which two approaches are available. In the traditional method the channel charge is calculated from the product of the dielectric capacitance and the gate overdrive ($Q_{\mathrm{ch}} = C_{\mathrm{ox}}(V_{\mathrm{g}} - V_{\mathrm{th}})$), which is applicable for device operation in the above-threshold, drift-limited regime. However, there are a couple of deficiencies associated with this technique which would lead to an inaccurate mobility calculation, the first being the calculation is only an approximation of the actual charge that resides in the channel. The assumption of a linear relationship between Q_{ch} and V_{g} is not particularly accurate near threshold. In addition, C_{ox} does not necessarily solely represent the dielectric capacitance. While this is a fair approximation for MOSFETs with thick gate dielectrics, for thin dielectrics C_{ox} needs to take into account the series capacitances associated with the effects of charge quantisation and DOS [99], both of which may not be strictly known. Secondly, the accuracy of the method largely depends on the accurate extraction of the threshold voltage.

A more accurate approach for determining the channel charge density is based on the split C–V technique, first proposed by Koomen [171] and later adopted by Sodini *et al.* [172] in the extraction of carrier mobility. In this method, Q_{ch} is obtained directly from the measurement of the gate-to-channel capacitance (C_{gc}) in the following way:

$$Q_{\mathrm{ch}} = \frac{1}{W_{\mathrm{g}}\, L_{\mathrm{g}}} \int_{-\infty}^{V_{\mathrm{g}}} C_{\mathrm{gc}}\,(\,V_{\mathrm{g}}\,)\,\mathrm{d}V\,. \tag{4.25}$$

The split C–V method has become a standard for mobility extraction in MOSFETs due to its simplicity and marked improvement over the traditional technique for Q_{ch} determination. However, the method is still open to a few sources of error that could potentially affect its accuracy [147]. One source of error is related to the source/drain series resistance which as previously highlighted acts to reduce the drain bias applied

across the device channel (see Equation (2.49)). To correct for this, V_{d} in Equation (4.24) should be replaced with $(V_{\mathrm{d}} - R_{\mathrm{sd}}.I_{\mathrm{d}})$.

Another common source of error arises from the effect of interface traps on the capacitance measurement. The additional capacitive contribution of interface traps (C_{it}) to the measured C_{gc} would result in the overestimation of the channel charge. As a consequence the effective mobility is underestimated. To minimise the ac trap response, the capacitance measurement should be performed at a sufficiently high frequency, typically in the range of 100 kHz to 1 MHz.

A third source of error is introduced from the C–V and I–V measurements being obtained at different bias conditions. While C_{gc}, and therefore Q_{ch}, are obtained at a drain bias of 0 V, a small, non-zero drain bias is applied during the I_{d}–V_{g} measurement. This introduces an error in the mobility data close to $V_{\mathrm{g}} = V_{\mathrm{th}}$. Ideally the smallest possible drain bias should be used for the drain current measurement to minimise this error. The measurement however becomes noisy for too small a value. Hence a drain bias in the range of -20 mV to -50 mV is typically applied [147].

4.4.2.4 Interface Trap density

The subthreshold swing (SS), a measure of how efficiently MOSFETs switch between on- and off-state operation, also provide useful information on the quality of the dielectric/semiconductor interface as previously highlighted. Defined as the gate bias required to induce a one decade change in the drain current ($dV_{\mathrm{g}}/d(log_{10}I_{\mathrm{d}})$), the swing can also be expressed as [156]:

$$SS = \ln(10)\,\frac{kT}{q}\left(1 + \frac{C_{\mathrm{d}} + C_{\mathrm{it}}}{C_{\mathrm{ox}}}\right). \tag{4.26}$$

It can be observed that interface states have a direct impact on SS as the associated capacitance ($C_{\mathrm{it}} = q\,D_{\mathrm{it}}$) is in parallel with depletion layer capacitance (C_{d}). This implies that in the presence of a large interface trap density, the swing is drastically increased, resulting in the degradation of device performance. The flatband MOSFETs realised in this work are essentially quantum-well (QW) devices that operate in accumulation-mode, for which $C_d \ll C_{\mathrm{ox}}$. Accounting for this and rearranging Equation (4.26), the interface trap density can be approximated as:

$$D_{it} \approx \frac{C_{\mathrm{ox}}}{q}\left(\frac{q\,SS}{\ln(10)\,kT} - 1\right). \tag{4.27}$$

This method of determining D_{it} is referred to as the subthreshold slope technique. It should be noted that D_{it} obtained from Equation (4.27) is an effective interface trap density that includes contributions from defect states at the dielectric/semiconductor interface and within the oxide, in close proximity to the interface [173]. For an accurate D_{it} extraction the subthreshold swing should be derived from I–V measurements obtained on transistors with $L_{\mathrm{g}} \geq 20\,\mu\mathrm{m}$, to exclude SCE on the measured characteristics [174].

4.5 Metrology

The measurement equipment used in this work varied depending on the type of characterisation required of material, contacts, gate stack and the overall device. Hall characterisation for measuring the resistivity, carrier concentration and mobility of the semiconductor was performed using two different measurement setups depending on the temperature requirement. For room temperature measurements the Biorad HL5200 Hall measurement system, capable of on-wafer probing, was used. It comprised of a permanent magnet with a field of 0.32 T, and the probes were connected to a RH2035 switchbox to enable resistance measurements for all possible permutations of the sample geometry. For automated instrumentation control coupled with data acquisition both the Hall system and the switchbox were computer controlled.

Hall measurements, in the temperature range of 10 K to 300 K, were undertaken using the Hall rig configured by Dr. Gary Paterson, located in the Department of Physics, University of Glasgow. Several instrumentation are comprised within the measurement setup. The I–V characteristics of the sample are measured using a Keithley model 2602 source measurement unit. Switching between sample contact configuration during Hall measurements is provided by the Keithley model 7001 switchbox. An electromagnet, powered by a 8 A supply, provides a magnetic field of 0.58 T at the maximum current rating. Temperature variation or stabilisation is achieved by means of a closed circuit helium cryostat, a model 8200 Cryodyne compressor and a model 502 Oxford Intelligent Temperature Controller (ITC). The sample for measurement has to be mounted into a chip carrier. This is then fitted into a chip carrier socket on the sample holder located inside the cryostat. Ultrasonic bonding is used to make electrical connections between the electrodes of the sample and the chip carrier. A semiconductor diode mounted on the sample holder is used for sensing the temperature of the sample. Most of the instruments are connected to a controlling computer running Labview software to allow for remote and automated instrument control.

All AC and DC measurements required for contact, gate stack and device characterisation are performed on-wafer using a semi-automatic Cascade Microtech Summit 12000 probe station. Stage control to select a specific sample site and to switch between sites can be done manually or under automated control using the Cascade Microtech Nucleus software package running on a control computer. Wafer mapping capability, provided within the software, allows for the selection of multiple sites for measurement automation whereby the stage is stationary at each site during a measurement and then moves to the next site to run a repeat of the measurement. For the automation to work correctly the separation between neighbouring sites, based on the x-y coordinate system, must be identical in the wafer map. Such automation capability allows for rapid and uninterrupted measurements of samples with hundreds of devices. Measurement and data acquisition capability is provided by an Agilent B1500A semiconductor parameter analyser running the Agilent EasyEXPERT software. Measurements can be configured to run in several modes; single and repeated measurements of one test setup or a multitude of different setups. The B1500A is equipped with a modular mainframe comprising a number of high-power and high-resolution source/monitor units (SMUs) for I–V sweeps and a multi-frequency capacitance measurement unit (MFCMU) for C–V sweeps in the frequency range of 1 kHz–5 MHz. In addition, SMU CMU unify unit (SCUU)/guard switch unit (GSWU) enables the measurements to be switched between AC and DC accurately and effortlessly. Communication between the probe station, B1500A and control computer is achieved by means of a general purpose interface bus (GPIB) connection.

4.6 Chapter Summary

This chapter presented the various characterisation techniques used in evaluating the metrics related to the grown epitaxial material, ohmic contacts, gate stack and overall device. The description began with the Hall effect and van der Pauw techniques for extracting the material transport data, followed by an outline of the transmission line model for deriving contact resistances. Attention was drawn to the importance of obtaining accurate measurements and the guidelines for this were underlined. Moreover the significance of correct metric extraction and interpretation was emphasised. The focus then shifted to MOS device characterisation. Relevant parameters that needed to be extracted from measurements obtained on MOS capacitors and MOSFETs in this work were presented and the commonly used extraction methods were discussed. Where relevant limitations and errors associated with the methods were highlighted and possible mitigating solutions were outlined. The discussions aided in establishing the appropriate extraction methods

for adoption in this work. The chapter concluded with an overview of the measurement kit utilised for characterisation during the different stages of transistor development.

5

Heterostructure Engineering

5.1 Introduction

Two aspects of device performance need to be addressed in the design of the MOSFET layer structure, the first being the maximisation of the drive current. For this, the channel transport needs to be optimised for charge carriers of both a low effective mass and a large concentration. Lower effective mass provides for a higher carrier mobility in long-channel devices, whilst increasing the injection velocity in ballistic channels [14]. High carrier concentrations can be obtained by increasing the doping in the device. The fundamental difficulty of realising a complementary logic solution using III-V compound semiconductors, as previously emphasised, is the high effective mass and thus low mobility of holes in these materials. A technological approach to enhance the hole mobility of the channel is the introduction of strain. Up to 200% enhancement in hole mobility has been obtained in Si CMOS transistors featuring uniaxially compressively strained channels [175]. Strain engineering is also a promising approach for enhancing the hole mobility in III-Vs. Theoretical results indicate a 2-4× improvement in the hole mobility of III-V semiconductors (GaAs, $In_{0.53}Ga_{0.47}As$, GaSb and InSb) from the incorporation of 2% biaxial compressive strain in the channel [176]. Practically, this is accomplished with heterostructures in which dissimilar materials are stacked on top of each other, using epitaxial growth techniques such as MBE [135]. Through judicious selection of materials and compositions, the heterostructure can be tailored for quantum confinement, which augments mobility by reducing the hole effective mass and/or the scattering [177], and strain, both of which are advantageous for enhancing channel transport. Biaxial compressively strained channels are realised through pseudomorphic growth on materials with a smaller lattice constant and the carrier transport is restricted to the channel layer

by cladding it with wider bandgap barrier layers. The large possibilities of material options available to III-V systems offer greater flexibility for engineering the heterostructure.

Apart from a high drain current, the additional requirement of the layer structure is to maintain electrostatic gate control of the channel, of importance to the device metrics of subthreshold swing, transconductance and DIBL. By adopting an appropriate device architecture that minimises the gate to channel separation and thus the CET, the electrostatic integrity of the device can be maximised [178].

In this chapter, the techniques of bandgap and strain engineering, collectively heterostructure engineering, are investigated for the optimisation of transport properties in III-V p-channel MOSFETs. The appropriate device architecture and the design of the III-V material layer stack are first discussed. Following this, the growth of the material layers used in this work are briefly mentioned. The test structures used in evaluating the transport performance of the heterostructures grown for the study, along with the fabrication and characterisation methods used, are further described. Thereafter, a systematic exploration of parameter space in the design of the heterostructure to optimise channel transport is presented. The experimental results from the study are reported and discussed.

5.2 Device Architecture and Layer Design

A number of planar device architectures have been investigated for III-V MOSFET development as illustrated in Figure 5.1 [179]. In the inversion-mode architecture (Figure 5.1(a)), a III-V equivalent of a traditional bulk Si MOSFET, the inversion channel is formed at the gate dielectric/semiconductor interface and the S/D regions are created by ion implantation. A variant of the inversion-mode MOSFET is the regrown source-drain architecture shown in Figure 5.1(b). This architecture comprises highly doped, regrown S/D regions to mitigate the access resistance problem resulting from the limitations of ion implantation in III-Vs. From a mobility enhancement perspective, the inversion-mode and regrown source-drain architectures are not ideal for device implementation. The greater impact of ionised impurity scattering on charge carriers, from the presence of dopants in the channel, would compromise channel mobility in the inversion-mode architecture. This is circumvented in the regrown source-drain architecture which employs an undoped channel. Moreover, carrier confinement akin to that achieved in ultra-thin body devices is obtained via the inclusion of a wider bandgap material underneath the channel. Despite these advantages, the channel of the regrown architecture is formed

at the dielectric/semiconductor interface, which is also the case for the inversion-mode architecture. The close proximity of the channel to the gate dielectric in the surface-channel configuration makes the channel carriers more susceptible to interface-related Coulomb scattering, interface roughness and dielectric-related scattering, resulting in mobility degradation [81, 92, 180, 181]. The impact of these scattering effects on mobility can be reduced with a buried-channel configuration in which the channel is spatially separated from the dielectric by a wider bandgap material [81, 92, 180, 181]. This is adopted in both the flatband-mode (Figure 5.1(c)) and recess gate (Figure 5.1(d)) architectures. Similar to the regrown architecture, the channels are also clad by wider bandgap materials providing for charge control comparable to ultra-thin body silicon-on-insulator (SOI) MOSFETs. The enhanced mobility in conjunction with the electrostatic confinement benefits render both the flatband-mode and recess gate architectures favourable for device implementation. Shown in Figure 5.2 are the experimental room temperature electron mobilities of $In_xGa_{1-x}As$ channels measured as a function of sheet density for the various

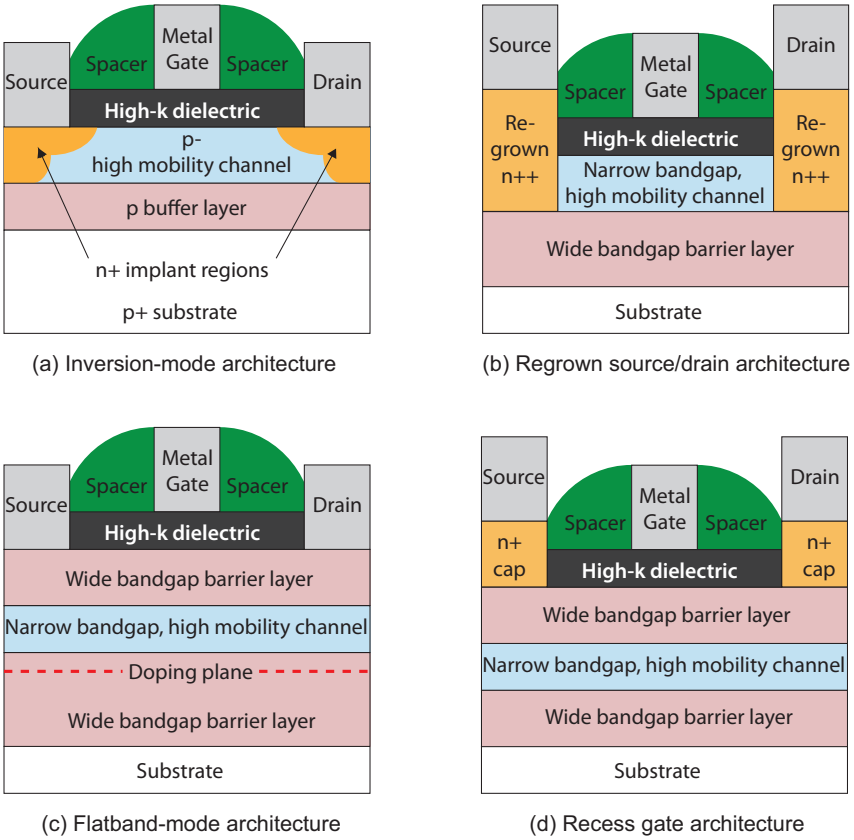


Figure 5.1: Various planar device architectures under investigation for III-V MOSFET development: (a) inversion-mode, (b) regrown source/drain, (c) flatband-mode and (d) recess gate architectures [179].

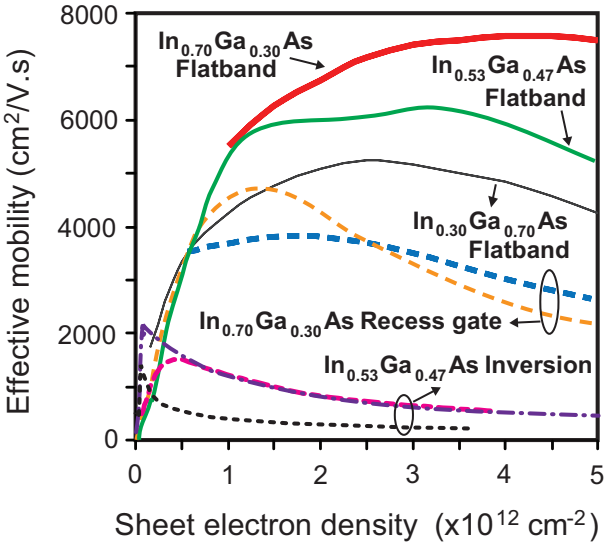


Figure 5.2: Room temperature electron mobilities of various $In_xGa_{1-x}As$ channels incorporated in devices based on the flatband-mode [179, 180, 182], the recess gate [181, 183] and the inversion-mode [184–186] architectures.

device architectures. It is notable that the devices based on the flatband-mode architecture demonstrate higher effective mobilities compared to the recess gate architecture. Based on this evidence, the flatband-mode architecture was chosen for III-V p-MOSFET development in this project.

Having established the device architecture for investigation, the next step is to tailor the design of the heterostructure for enhanced channel transport. Drawing on the review presented in Section 2.4.2, the salient parameters that influence the QW channel transport are highlighted in Figure 5.3. The introduction of biaxial compressive strain in the channel leads to energy splitting between the heavy-hole (HH) and light-hole (LH) bands ($\Delta_{\text{HH-LH}}$), with the HH being higher in energy. As a result, more holes will first occupy the HH band before starting to fill the LH band. Due to the lighter in-plane effective mass of the HH band more holes would have a lower effective mass and thus the effective hole mobility would be higher. The HH-LH splitting also reduces interband scattering leading to a higher hole mobility. By increasing $\Delta_{\text{HH-LH}}$, through increased strain, the effective mass and scattering can be further reduced, leading to higher hole mobility. Additionally, valence band offset (VBO) may also be affected by strain due to changes in the bandgap. In the case of $In_xGa_{1-x}As$ (channel)/$In_{0.52}Al_{0.48}As$ (barrier), the band offset can be increased from 0.22 eV to 0.38 eV by changing the indium content of the channel from 0.53 to 1 [73]. It is important to note that with the increase of strain, the QW channel thickness needs to be maintained below the critical layer thickness (CLT) to prevent strain

relaxation leading to dislocation generation; CLT decreases with increasing strain.

Quantum confinement imposed by the reduction of the channel thickness (t_{chan}) with increasing strain is of further advantage as this induces subband splitting as shown in Figure 5.3. As before, the majority of holes will first occupy the highest energy subband, HH1, with the lower in-plane effective mass, before starting to populate subbands that are lower in energy and possess heavier in-plane mass (HH2 or LH1). Similar to strain, quantum confinement results in both the effective mass and scattering (interband and intersubband) to be reduced, leading to a higher hole mobility. The mobility can be raised further by thinning the channel, which acts to strengthen the confinement for increased subband splitting. However, too narrow a channel can also be detrimental to channel transport as the effects of hetero-interface scattering becomes enhanced [73, 76, 177].

The maximum number of holes confined in the QW channel is determined by the VBO at the hetero-interface between the channel and the wider bandgap barrier layers, the spacer layer thickness (t_{spa}) and the doping concentration. For good hole confinement, it is important to choose a barrier material that can provide a large enough VBO [177].

The adoption of modulation doping in which the dopant atoms are located in the wider bandgap barrier/buffer layers as opposed to the channel itself minimises the effect of ionised impurity scattering on holes occupying the channel, thereby mitigating mobility degradation. The inclusion of a spacer layer between the channel and the doping layer results in additional mobility enhancement since the scattering effect is further suppressed.

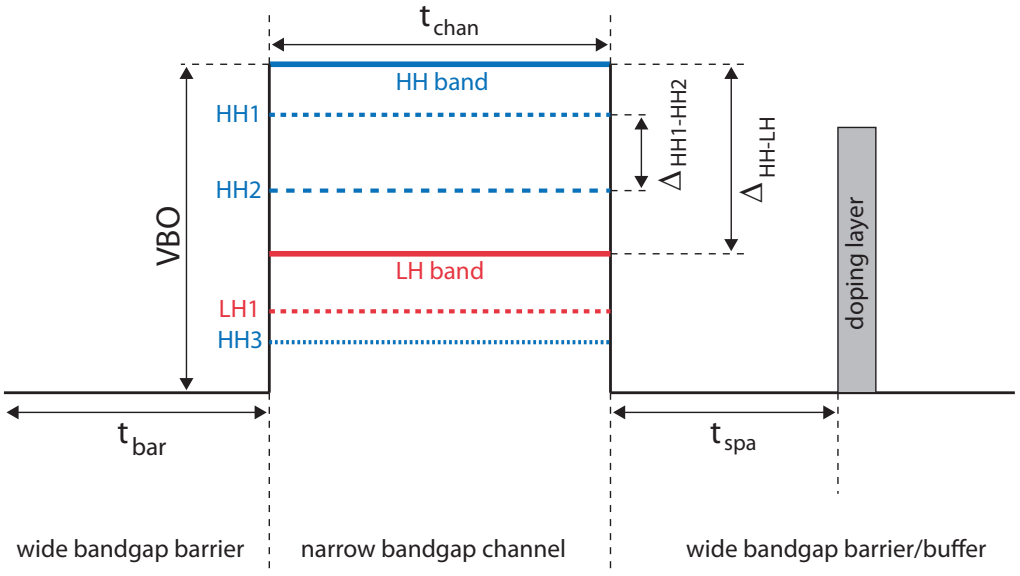


Figure 5.3: A simplified valence band diagram of a heterostructure with key parameters impacting QW channel transport identified.

By increasing the thickness of the spacer, the mobility is improved further. However, the tradeoff with employing a thick spacer is the reduction of charge transfer from the doped layer to the QW channel leading to a decrease in hole density [187, 188]. This could also result in parallel conduction in the doping layer [187]. It is therefore important a spacer of a suitable thickness is used to realise a channel of high mobility and large hole concentration. The efficiency of charge transfer can be improved further by using a delta-doped layer rather than a uniformly doped layer [189]. Furthermore, more holes will populate the channel from increasing the impurity concentration of the doping layer. However, too high a concentration could also give rise to the doping layer acting as a parallel conducting layer.

Another key parameter in the design of the heterostructure is the thickness of the barrier (t_{bar}) between the channel and the gate dielectric (see Figure 5.1(c)). A thick barrier is desirable as this minimises interface- and dielectric-related carrier scattering, thereby alleviating mobility degradation. However, this increases the CET which in turn degrades the electrostatic gate control of the channel [81, 92].

In the III-V material system, a number of channel/barrier material options are available for engineering the heterostructure for quantum confinement and strain effects. These options can be categorised as arsenide-based and antimonide-based. The arsenide-based options are (In)GaAs/(Al)GaAs, $\mathrm{In}_{x\geq 0.53}\mathrm{GaAs}$/InAlAs and $\mathrm{In}_{x\geq 0.53}\mathrm{GaAs}$/InP. The earliest reports of III-V p-channel heterostructure field-effect-transistor (HFET) were based on a GaAs/AlGaAs material option, which resulted in room temperature (RT) mobilities of 170–190 cm^2/Vs [107, 190, 191]. In an effort to improve transport performance, subsequent p-channel work utilised $\mathrm{In}_x\mathrm{Ga}_{1-x}\mathrm{As}$ channels with either AlGaAs or GaAs barriers for x $\leq$ 0.35 [187, 192–200] and InAlAs or InP barriers with $0.53 \leq x \leq 0.95$ [73, 201, 202]. The various works yielded RT hole mobilities of 191–390 cm^2/Vs. Of these material options, channels based on $\mathrm{In}_{x\geq 0.53}\mathrm{GaAs}$ are appropriate for investigation due to the lower hole effective mass that can be obtained with increasing In mole fraction, x. In comparing the two options, the $\mathrm{In}_{x\geq 0.53}\mathrm{GaAs}$/InP material combination is favourable for its VBO. For an $\mathrm{In}_x\mathrm{Ga}_{1-x}\mathrm{As}$ channel in the range $0.53 \leq x \leq 1$, the InP barrier provides a larger band offset of 0.43–0.7 eV compared to the offset of 0.22–0.38 eV provided by the InAlAs barrier [73]. As a result carrier confinement is enhanced for the $\mathrm{In}_{x\geq 0.53}\mathrm{GaAs}$/InP system. However, it is not possible to grow phosphorous-containing materials in the MBE system located at the University of Glasgow. For this reason, the arsenide-based III-V p-channel MOSFET development in this project was restricted to the $\mathrm{In}_{x\geq 0.53}\mathrm{GaAs}$/InAlAs material option.

Given the bulk hole mobilities of antimony containing compound semiconductors (GaSb and InSb) are the largest of any III-V materials, it is only logical that antimonide-based materials are also considered for p-channel MOSFETs. Antimonides afford several material options for heterostructure engineering. One option is a GaSb channel confined by an $AlAs_{x\geq 0.1}Sb_{1-x}$ barrier. The large VBO of 0.46 eV obtained for GaSb lattice-matched to AlAsSb provides for good confinement of holes in the QW channel [203]. Strain in the QW channel is controlled by changing the barrier layer composition, which also enables the VBO to be increased further. The use of a binary semiconductor as the channel material is further advantageous since alloy scattering is avoided in the channel. RT hole mobilities in the range of 1200-1500 cm^2/Vs have been obtained in 5-10 nm thick QW channels subject to 1-3% biaxial compressive strains [74, 75]. Quantum wells of $In_xGa_{1-x}Sb$ clad by AlGaSb is another material option. Strain is incorporated by changing the indium composition of the InGaSb channel. For 5-12.5 nm thick channels subject to 1-2% compressive strains, RT hole mobilities in the range of 780-1500 cm^2/Vs have been realised [76, 204]. Although the channel layer alloy composition does not alter the band offset, good confinement of holes in the channel is still attainable due to the large VBO of 0.43 eV obtained with InGaSb/AlGaSb [205]. A final material option of $InSb/Al_xIn_{1-x}Sb$ is similar to the GaSb/AlAsSb option, whereby strain is controlled through the barrier layer composition. A RT hole mobility of 1230 cm^2/Vs was obtained for a 5 nm thick channel under 1.9% compressive strain [77]. Despite the fact that VBO can be controlled through the barrier layer composition, the band offsets tend to be lower compared to the other antimonide-based options. The band offset of $InSb/Al_{0.4}In_{0.6}Sb$, with the channel subject to 1.9% strain, is only 0.3 eV [177]. The enhanced BTBT leakage currents resulting from the small bandgap of InSb is also a concern. Taking into consideration the merits of the various antimonide-based material options, both GaSb/AlAsSb and InGaSb/AlGaSb appear suited for investigation. It is further favourable to employ a material combination that enables a complementary logic solution of high performance to be achieved. In this regard, the InGaSb/AlGaSb system is more advantageous since large band offsets for both electrons and holes can be achieved with strain effects. In the case of GaSb/AlAsSb, strain, increasing the VBO decreases the conduction band offset (CBO), resulting in the degradation of electron confinement in the channel. Therefore, in this work, antimonide-based p-channel MOSFETs are developed using $In_{x\geq 0.1}Ga_{1-x}Sb$ channels confined by AlSb barriers.

Since the barrier layer above the channel forms an intimate contact with the gate dielectric in the flatband MOSFET architecture (see Figure 5.2(b)), and therefore has

a direct impact on the dielectric/semiconductor interface properties, it is further important to evaluate the suitability of the topside barrier material for high-k dielectric integration. The presence of aluminium in the barrier layers of both the arsenide-based and antimonide-based material options is an issue since surface layers terminated with aluminium-containing materials are prone to rapid oxidation on exposure to air [181, 206–208]. Among the native oxides left behind on the surface from the oxidation, oxides of aluminium (AlO_x) are particularly difficult to passivate [206, 209]. This gives rise to interface traps, which screen the gate potential from modulating the channel charge. As a result, transistor performance is degraded. This can be avoided by terminating the surface with a thin cap layer that is aluminium-free, lattice-matched to the underlying barrier layer and on which good passivation can be realised for a high quality dielectric/III-V interface. Based on this criteria, $In_{0.53}Ga_{0.47}As$ was determined to be suitable as a cap layer for the arsenide-based material option while GaSb was favourable for the antimonide-based option.

A final comment is regarding the critical layer thickness of the channel. As previously highlighted, the increase of biaxial compressive strain in the channel must be accompanied by a reduction of the channel thickness below CLT. The dependence of the critical thickness on the alloy content, x, of the channel and the resulting strain for the chosen arsenide and antimonide-based material options, calculated using the People-Bean [109] and Matthews-Blakeslee [110] models, are shown in Figure 5.4. The People-Bean model is more accurate for lower strain ($< 1\%$) while the Matthews-Blakeslee model gives better

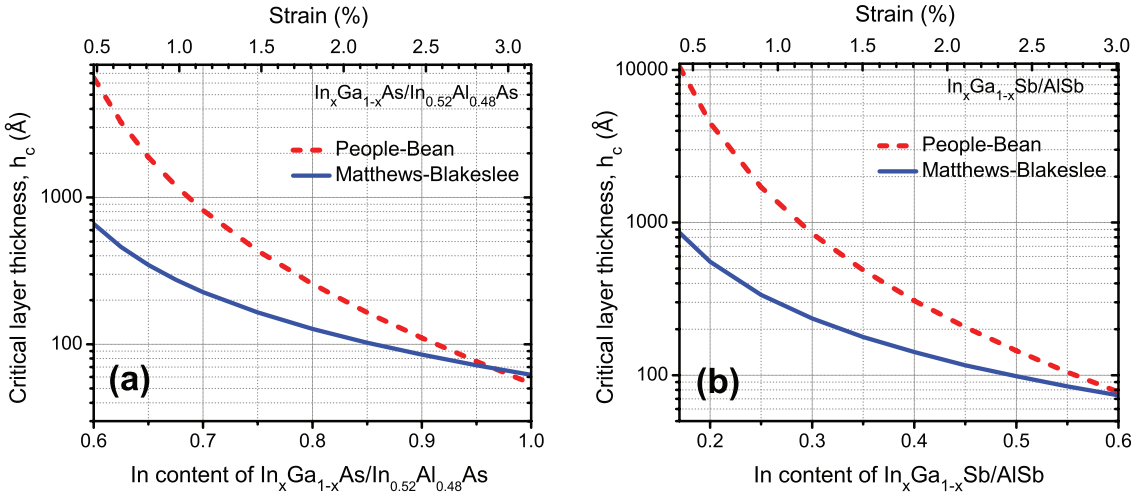


Figure 5.4: Critical layer thickness as a function of strain and In content for (a) $In_xGa_{1-x}As$ layer grown on $In_{0.52}Al_{0.48}As$ and (b) $In_xGa_{1-x}Sb$ layer grown on AlSb, estimated using People-Bean [109] and Matthews-Blakeslee [110] models.

results for higher strain [111].

Based on the aforementioned discussion, the cross-sections of the layer designs employed for arsenide- and antimonide-based flatband MOSFET development in this work are illustrated in Figure 5.5. The arsenide-based layers are grown on a semi-insulating (SI) (100) InP substrate and the antimonide-based layers are grown on a SI (100) GaAs substrate. A thick buffer layer (1 μm) is used in the antimonide-based design to accommodate the 8% lattice mismatch between AlSb and the GaAs substrate. A 2 nm cap/2 nm barrier is included between the gate dielectric and the channel in the layer designs based on both material options. Doping strategy is a key variant that is investigated as part of the layer design optimisation. As shown in Figure 5.5, the delta (δ)-doping plane, separated by a spacer layer, is located either above the channel or below the channel. The former strategy is referred to as non-inverted doping while the latter is known as inverted doping. It is worth noting that single-sided doping is employed for all layer designs investigated in this work. In addition to this, the spacer thickness, the channel thickness, the channel composition (strain), and the δ-doping density are the other design parameters investigated for channel transport optimisation.

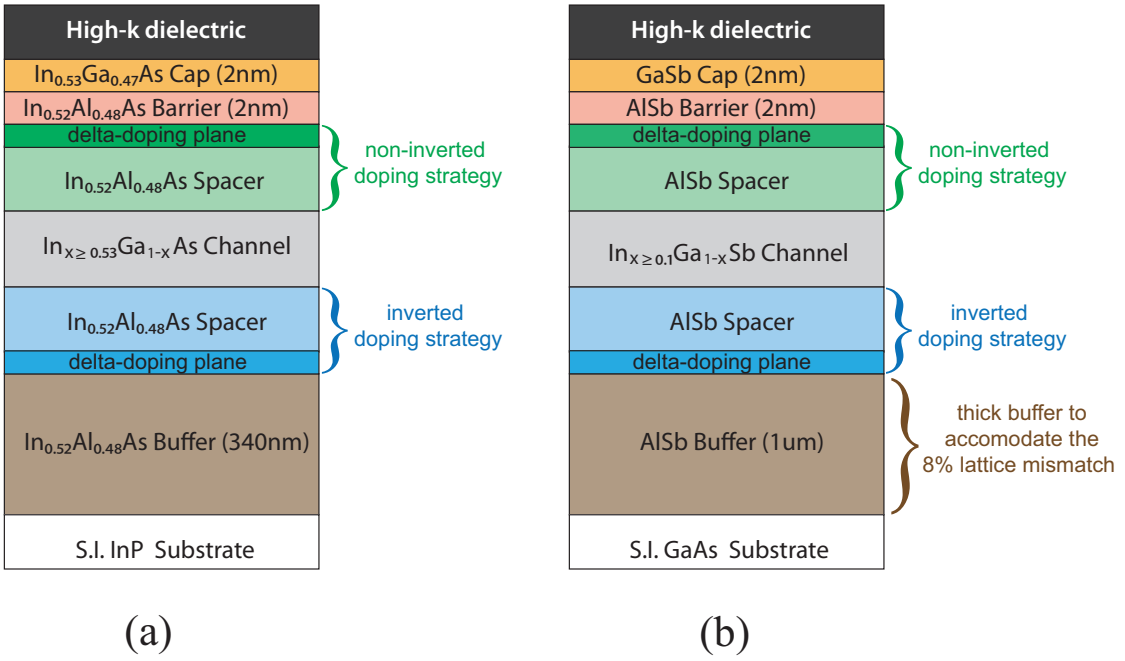


Figure 5.5: Cross-sections of (a) $In_{x\geq 0.53}Ga_{1-x}As/In_{0.52}Al_{0.48}As$ and (b) $In_{x\geq 0.1}Ga_{1-x}Sb/AlSb$ p-channel flatband MOSFET layer designs.

5.3 Growth, Fabrication and Measurement

At this stage of the project, suitable high-k dielectrics were still under development. As a result, transport in the arsenide- and antimonide-based layer designs were investigated using $In_{0.53}Ga_{0.47}As$ and GaSb terminated wafers of suitable structure to mimic the electrostatics of a III-V p-channel MOSFET. This architecture is referred to as pre-MOSFET as it shared the underlying structure with the flatband MOSFET design. All pre-MOSFET structures investigated as part of the layer design and optimisation study were grown by MBE; arsenide-based wafers were grown at the University of Glasgow and antimonide-based wafers were grown at Texas State University. Beryllium was used as the dopant impurity for all growths.

Two types of test structures were used for Hall characterisation of samples detailed in the following section unless otherwise stated. Room-temperature Hall measurements were carried out on Van der Pauw structures (see Figure 4.3(b)). Such structures were realised by depositing In/Zn metal pads ($150\,\mu m \times 150\,\mu m$) on four corners of a $15\,mm \times 15\,mm$ sample and subsequently annealing at 360 °C for 20 s to form ohmic contacts which could be measured using wafer probes. Temperature dependent transport measurements, on the other hand, were performed on a 6-contact Hallbar structure (see Figure 4.2). The Hallbar was designed following the geometrical guidelines highlighted in Section 4.2 and patterned using two levels of photolithography. In the first lithographic level, ohmic contacts were defined by the thermal evaporation of Au/Zn/Au metals followed by rapid thermal annealing (RTA) at 360 °C for 3 s. Mesa isolation was realised in the subsequent level by means of a wet chemical etch comprising orthophosphoric acid, hydrogen peroxide and water. Details of the process flow for Hallbar fabrication is listed in Appendix A.1. The metrology adopted for room temperature and temperature dependent Hall characterisation is outlined in Section 4.5.

5.4 Experimental Investigation

In this section, the parameter space investigated for the arsenide- and antimonide-based layer designs, using the pre-MOSFET structure, are presented. The transport results from the study are reported and discussed.

5.4.1 Arsenide-Based Layer Designs

5.4.1.1 Lattice-Matched Design

For the purpose of assessing the impact of strain on the channel transport properties, a lattice-matched III-V materials "testbed" first needed to be established. As such, the first-generation p-channel pre-MOSFETs grown were based on a lattice-matched, non-inverted design architecture as shown in Figure 5.6. The structure comprised an $In_{0.53}Ga_{0.47}As$ QW channel of a fixed thickness of 10 nm. A thickness of 4 nm and a bulk doping density of $2.5\times10^{18}\,cm^{-3}$ ensured that the cap layer was depleted. The parameter spacing investigated for the lattice-matched structures included the δ-doping density, spacer thickness and barrier thickness, with the channel-to-surface distance maintained at 30 nm.

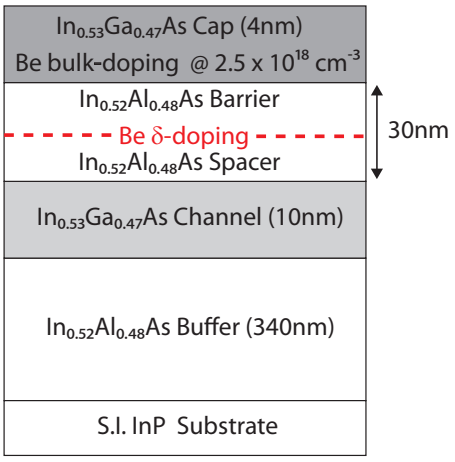


Figure 5.6: Cross-section of a typical lattice-matched, non-inverted arsenide-based pre-MOSFET.

An overview of the wafers grown as part of this investigation, together with the hole mobility (μ_h), sheet concentration (p_s) and sheet resistance (R_{sh}) associated with each growth are illustrated in Table 5.1. The first two wafers grown, c545 and c550, featured a 5 nm spacer with a corresponding barrier thickness of 25 nm, however c550 was grown

Table 5.1: Summary of structural/growth conditions and room temperature Hall data of lattice-matched, non-inverted arsenide-based pre-MOSFETs.

Wafer	Barrier thickness (nm)	δ-doping (cm^{-2})	Spacer thickness (nm)	Growth temperature (°C)	μ_h (cm^2/Vs)	p_s (cm^{-2})	R_{sh} (Ω/sq)
c545	25	4×10^{12}	5	450	62	2.4×10^{12}	38796
c550	25	2.85×10^{12}	5	450	67	1.8×10^{12}	47791
c567	23	2.85×10^{12}	7	400	65	1.6×10^{12}	57938

with a lower δ-doping density. Based on the transport measurements comparable hole mobilities were obtained for both wafers. On the other hand the higher sheet concentration measured for c545 in comparison to c550 is a direct result of the larger δ-doping density in the former, resulting in more charge transfer into the QW channel. Although the measured mobilities of these wafers are in close agreement with the experimental mobility value of $55\,cm^2/Vs$ reported for a lattice-matched, non-inverted InGaAs QW channel structure [73], the obtained values are rather modest compared to a mobility of $145\,cm^2/Vs$ derived from simulations of a relaxed InGaAs channel at a sheet concentration of $2\times10^{12}\,cm^{-2}$ [176]. A plausible explanation that would account for the modest values is the migration of dopants from the δ-doping plane into the channel as Be dopants tend to out diffuse more readily in aluminum containing alloys [210]. The extent of dopant diffusion in wafers c545 and c550 was estimated to be about 10.5 nm based on their substrate growth temperature of 450 °C [211]. As such it is possible the dopants have diffused past the spacer into the channel as the diffusion length is twice as big as the spacer thickness. The dominance of ionised impurity scattering on channel carriers would account for the modest mobilities in these wafers. Dopant diffusion can be minimised by lowering the growth temperature in conjunction with increasing the spacer thickness. For a growth temperature of 400 °C the diffusion length was estimated to be about 6 nm [211]. Taking this into consideration, wafer c567 was grown with a 7 nm thick spacer at a growth temperature of 400 °C. However this did not result in any improvement as the measured mobility of wafer c567 was again comparable to that obtained from wafers c545 and c550. This seemed to suggest that dopant diffusion may not necessarily be the cause of the modest mobility of these wafers.

Alternatively the presence of two or more parallel conducting channels could account for the modest mobilities obtained from the lattice-matched wafers. Such additional parallel paths can arise from a number of effects including multiple subband occupancy

of a single 2DHG, presence of several spatially separated 2DHG within the QW or in the heterostructure, and free carriers in un-depleted regions of the δ-doped plane [212]. In such cases the data obtained from the Hall measurements are complex averages of the sheet concentration and mobility associated with each of the conducting channels and thus not necessarily representative of single channel transport [213]. Therefore it would be incorrect to assume that the measured Hall data is associated solely with single channel conduction. The presence of additional channels results in the underestimation and overestimation of the measured mobility and sheet concentration data, respectively [213].

To elucidate the presence of parallel conduction, temperature dependent Hall measurements were undertaken on wafer c545 and the results are displayed in Figure 5.7. The carrier mobility is observed to increase monotonically as the temperature is lowered from 300 K, reaching a value of 429 cm^2/Vs at 10 K. This is accompanied by a monotonic decrease in sheet carrier concentration. It is notable that characteristics usually associated with modulation doped structures, whereby the carrier density initially decreases with temperature and then becomes independent of temperature over a wide temperature window, is not observed for wafer c545. This absence of saturation in the sheet concentration with temperature variance is attributed to parallel conduction effects [214]. Apart from confirming the presence of parallel conduction, the temperature dependent Hall data provides no further information regarding the nature of the additional conduction paths

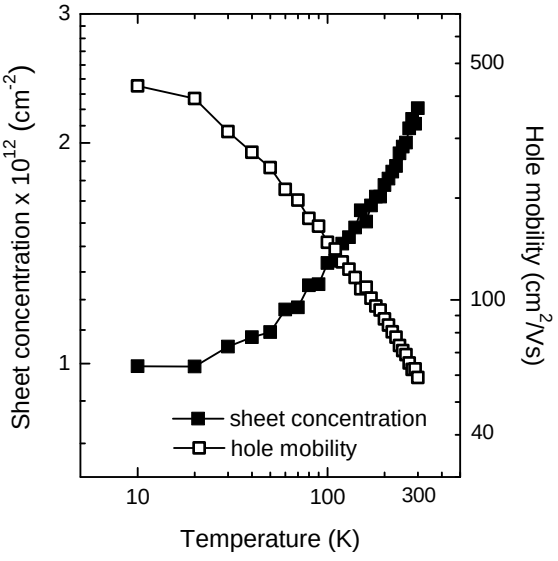


Figure 5.7: Hole mobility and sheet concentration as a function of temperature for wafer c545.

present in the lattice-matched wafers. However a few of the previously mentioned effects contributing to parallel conduction can be excluded based on the heterostructure design. The QW of modulation doped structures featuring single-sided doping will be asymmetric in shape. As such, the presence of more than one spatially separated 2DHG in the channel of the lattice-matched wafers is unlikely [215]. Additionally the cap has been designed such that is it depleted of carriers and hence will not contribute to parallel conduction either.

5.4.1.2 Strained Channel, Non-Inverted Design

The second-generation pre-MOSFETs grown were based on a non-inverted architecture featuring strained InGaAs QW channels as illustrated in Figure 5.8. A cap layer design similar to that utilised in the lattice-matched pre-MOSFET was employed. The barrier and spacer thickness were kept constant at 25 nm and 5 nm respectively. The structural/-growth parameters varied for the strained pre-MOSFETs are the δ-doping density, the In mole fraction, x, in the $In_xGa_{1-x}As$ QW and its equivalent strain, and the thickness of the QW channel.

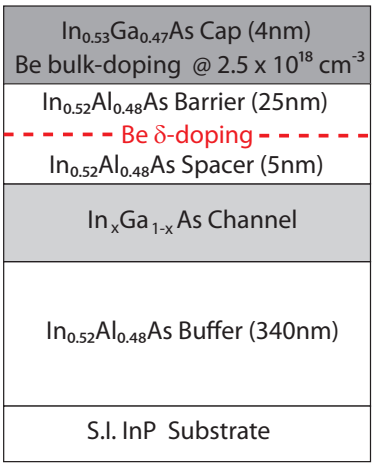


Figure 5.8: Cross-section of a typical strained, non-inverted arsenide-based pre-MOSFET.

A summary of the growth/structural conditions investigated for the strained, non-inverted pre-MOSFET structures and the associated room temperature material transport parameters are shown in Table 5.2. The first wafer grown, c646, was identical to wafer c550 with the exception of strain; wafer c646 featured a $In_{0.75}Ga_{0.25}As$ channel resulting in the QW being under 1.4% biaxial compressive strain. The first thing to note is that

Table 5.2: Summary of structural/growth conditions and room temperature Hall data of strained, non-inverted arsenide-based pre-MOSFETs.

Wafer	InAs mole fraction, x, of $In_xGa_{1-x}As$ QW (%)	Strain (%)	QW thickness (nm)	δ-doping (cm^{-2})	μ_h (cm^2/Vs)	p_s (cm^{-2})	R_{sh} (Ω/sq)
c646	75	1.4	10	2.85×10^{12}	336	1.19×10^{12}	15600
c665	85	2.1	7.5	2.85×10^{12}	433	1.33×10^{12}	10900
c674	85	2.1	7.5	4×10^{12}	427	1.6×10^{12}	9181

the mobility of wafer c646 is 336 cm^2/Vs. This is a five-fold increase compared to wafer c550. It is also interesting to note that the sheet concentration of wafer c646 is 33% lower compared to wafer c550. The larger valence band offset between the strained InGaAs channel and $In_{0.52}Al_{0.48}As$ spacer/barrier layers as opposed to the lattice-matched channel of wafer c550 should in fact result in more carriers populating the QW of wafer c646. On the contrary, the experimental data shows less carriers and the most likely explanation for this is parallel conduction in c550 due to un-depleted regions of the delta plane owing to the smaller valence band offset of 0.22 eV, which would account for the poor confinement of holes in the channel. If ionised impurity scattering due to the mechanism of dopant diffusion was limiting the hole mobility, as was speculated to be the case for the lattice-matched pre-MOSFETs, this should also persist in strained pre-MOSFETs. However, on the basis of the mobility obtained for wafer c646 it can be concluded that the growth of both the lattice-matched and strained structures are not likely affected by dopant diffusion. In order to enhance the transport properties, the InAs mole fraction of the QW channel was increased to 85%, with an equivalent strain of 2.1%, and the channel thickness was reduced to 7.5 nm in the next growth, wafer c665. The larger strain and enhanced quantum confinement of wafer c665 as compared to c646 resulted in a 30% and 12% improvement in the hole mobility and sheet concentration respectively. In an effort to enhance hole accumulation in the QW, the δ-doping density was increased to 4×10^{12} cm^{-2} in the subsequent wafer that was grown, c674, with all other structural/-growth parameters being identical to wafer c665. This resulted in a 20% improvement in the sheet concentration of wafer c674 compared to wafer c665, while the mobility only decreased by a margin of 1.5%.

5.4.1.3 Strained Channel, Inverted Design

An additional CET of 1.57 nm is incurred in the previously investigated non-inverted structures due the presence of the 5 nm $In_{0.52}Al_{0.48}As$ spacer layer separating the δ-doping plane from the QW channel. This can be avoided by moving the delta-doping plane below the channel, which in turn would improve the gate control of the channel. Therefore the third-generation pre-MOSFETs grown were based around strained InGaAs QW channels employing an inverted architecture as shown in Figure 5.9. The cap layer was identical to the non-inverted structures and was separated from the QW channel by a 25 nm thick $In_{0.52}Al_{0.48}As$ barrier layer. Since the most promising transport data for the non-inverted structure was obtained for an $In_xGa_{1-x}As$ QW channel with an In mole fraction of 85% and a thickness of 7.5 nm, a similar channel design was also adopted for the inverted structure. A 4 nm thick $In_{0.52}Al_{0.48}As$ spacer layer is interposed between the channel and the δ-doping plane. The structural/growth parameter spacing investigated for the inverted structures comprised the δ-doping density and the growth temperature.

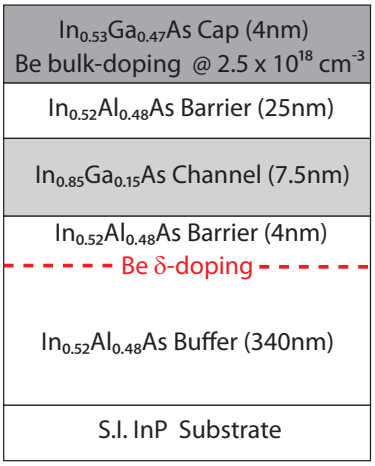


Figure 5.9: Cross-section of a typical strained, inverted arsenide-based pre-MOSFET.

An overview of the structural/growth parameters examined for the strained, inverted pre-MOSFETs together with the room temperature transport data are illustrated in Table 5.3. The first wafer, c726, grown featured a δ-doping density of $4\times10^{12}\,cm^{-2}$, resulting in a sheet concentration of $2.3\times10^{12}\,cm^{-2}$. However the mobility was substantially lower showing a 46% reduction in comparison to wafer c674, a non-inverted equivalent of wafer c726. The often lower mobilities resulting from inverted as opposed to non-inverted modulation doped structures is a well known phenomenon. A root cause of mobility degradation in inverted structures is dopant segregation, whereby during the growth process dopant

Table 5.3: Summary of structural/growth conditions and room temperature Hall data of strained, inverted arsenide-based pre-MOSFETs.

Material	δ-doping (cm^{-2})	Growth temperature (°C)	μ_h (cm^2/Vs)	p_s (cm^{-2})	R_{sh} (Ω/sq)
c726	4×10^{12}	450	230	2.3×10^{12}	11600
c734	3×10^{12}	400	318	0.97×10^{12}	20300
c735	4×10^{12}	400	306	1.37×10^{12}	15000

atoms preferentially spread (segregate) towards the direction of growth [216, 217]. As a result, the doping profile instead of being abrupt becomes broadened with a long tail extending towards the surface. In non-inverted structures this poses no problem since the channel is below the δ-doping plane. However in the case of inverted structures, the placement of the doping below the channel results in dopants segregating into the channel during the growth. The enhanced effect of ionised impurity scattering results in the mobility being degraded. This therefore accounts for the mobility value of 230 cm^2/Vs obtained for wafer c726.

A number of techniques such as growth interrupts [216, 217], thicker spacers [216] and lower growth temperatures [218, 219] can be employed to alleviate dopant segregation. The issue with the adoption of thick spacers is the reduction in carrier population in the QW channel. Also, growth interrupts were already utilised for all structures grown in this work. Therefore, for the next wafer, c734, the temperature was lowered from 450 °C from the previous growth to 400 °C. Additionally in an effort to suppress the number of dopants segregating into the channel, the doping was decreased to 3×10^{12} cm^{-2}. The resultant mobility of 318 cm^2/Vs of wafer c734 was reflective of the structural/growth modifications, showing an improvement of 38% over the mobility value of wafer c726. This indicates that dopant segregation has been suppressed due to the reduction in temperature. On the other hand the sheet concentration was measured to be just under 1×10^{12} cm^{-2} and is a result of the lower doping level. To determine if a mobility value similar to that of wafer c734 can be obtained at a higher doping level of 4×10^{12} cm^{-2} under the same growth temperature of 400 °C, a final wafer, c735, was grown. From the measured transport properties it is noted that the mobility of wafer c735 is comparable to that of wafer c734, with only a 3% difference between the two growths. It is therefore deduced that the key factor in suppressing dopant segregation is the growth temperature with the δ-doping density having minimal effect. The higher doping level further resulted in more carriers populating

the channel, with wafer c735 showing a 41% increase in sheet concentration from that of wafer c734. To confirm the hypothesis of dopant segregation and its suppression following a reduction in the substrate growth temperature, secondary ion mass spectroscopy (SIMS) measurements were undertaken on wafers c726 and c735. The resulting Be depth profiles are shown in Figure 5.10. For both wafers the Be concentration peaks at the δ-doped position are clearly distinct and show a similar value of approximately $2\times10^{19}\,\mathrm{cm}^{-3}$. While the Be profile shows a sharp turn on in both wafers, the turn off profile is distinctly different between the wafers. For wafer c726 the Be peak is accompanied by a distinct exponential tail extending towards the sample surface, indicative of dopant segregation. The Be peak of wafer c735, however, has a sharp turn off profile. This confirms the suppression of dopant segregation from the reduction in growth temperature.

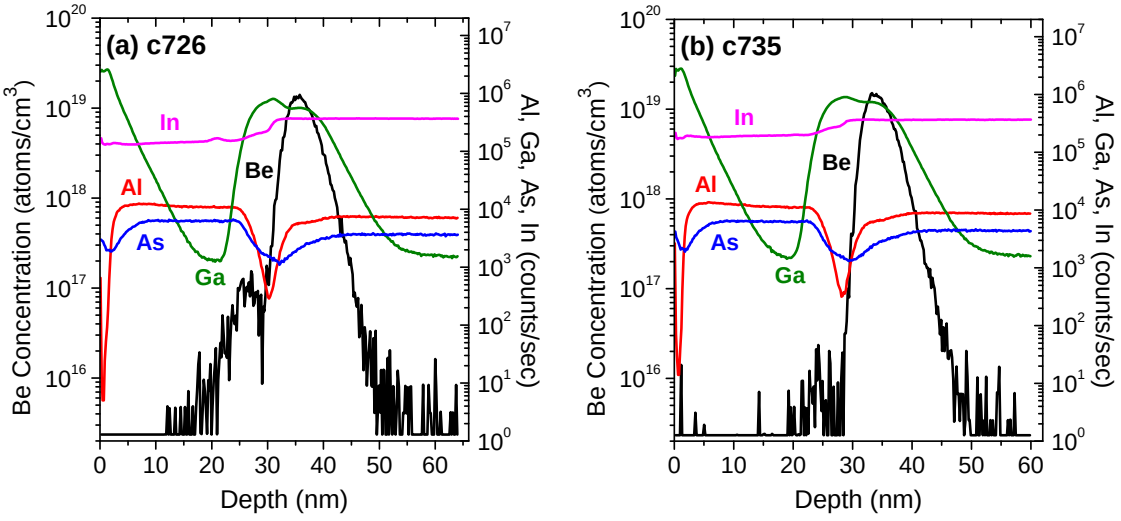


Figure 5.10: Experimental secondary-ion mass spectroscopy (SIMS) depth profiles of Be in inverted arsenide-based pre-MOSFETs: (a) c726 and (b) c735 wafers.

5.4.1.4 Performance Benchmark

In Figure 5.11, room temperature hole mobility as a function of channel sheet density is plotted for strained arsenide-based pre-MOSFETs investigated in this work. For comparison, the mobility data of strained InGaAs QWs from literature [73, 187, 192, 193, 195–199, 201, 202] are also included in the plot. It is notable that the non-inverted structures, subject to 2.1% biaxial compressive strain, demonstrate the highest hole mobilities ever reported for an InGaAs QW channel. The mobility data of the inverted structures also appear to be comparable to the best reported mobilities in literature.

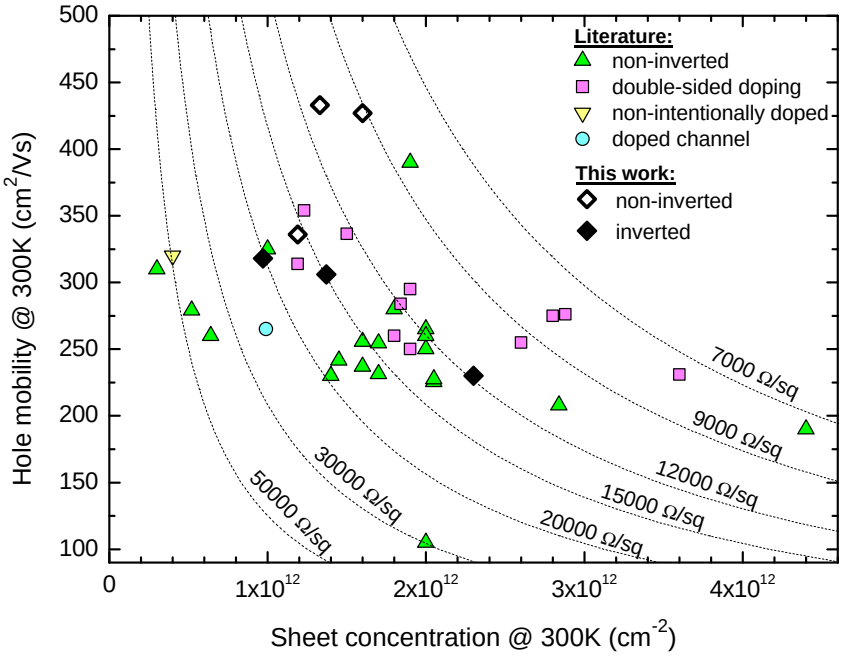
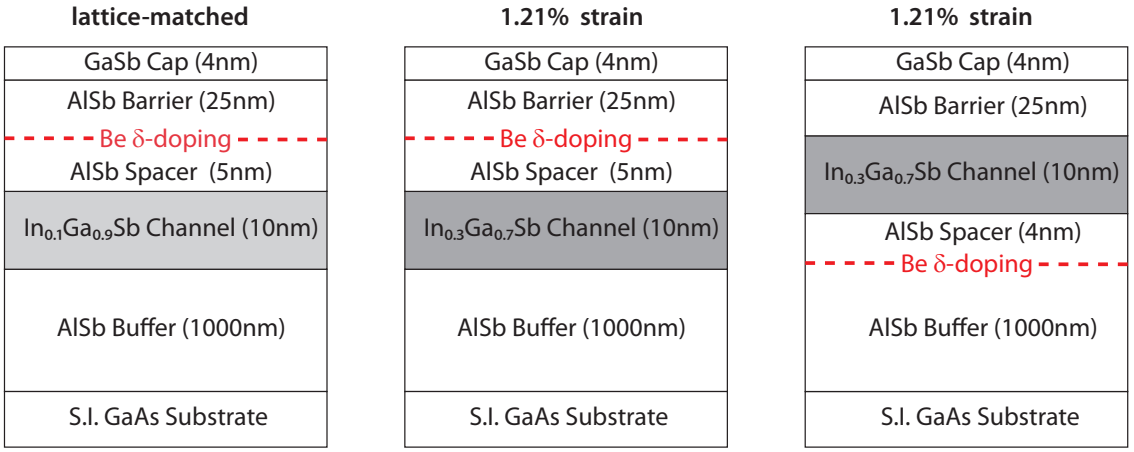


Figure 5.11: Room-temperature hole mobility as a function of sheet concentration for strained, inverted and non-inverted, InGaAs quantum wells investigated in this work. Data from literature for strained InGaAs quantum wells are also included for comparison [73, 187, 192, 193, 195–199, 201, 202].

5.4.2 Antimonide-Based Layer Designs

Access to MBE growth capability of antimonide-based materials only became available at a late stage in the project. Due to time constraints, a systematic investigation of parameter spacing, similar to that undertaken for the arsenide-based designs, was not carried out for the antimonide-based layer design. Shown in Figure 5.12 are three antimonide-based pre-MOSFETs grown for a quick evaluation of the channel transport properties. Common to all three pre-MOSFETs are a 4 nm GaSb cap, a 25 nm AlSb barrier, a 1 μm AlSb buffer, a 10 nm channel and 4×10^{12} cm^{-2} δ-doping. Doping strategy and channel strain are the only two variants between the growths. Wafers 7-133 and 7-136 are based on a non-inverted design while an inverted design is used for wafer 7-137. A lattice-matched QW channel is used in wafer 7-133. On the other hand, wafers 7-136 and 7-137 comprise a $In_{0.3}Ga_{0.7}Sb$ channel, resulting in the QW being under 1.21% biaxial compressive strain.

Contactless Hall measurements, using a Lehighton model 1605 measurement system, were carried out on the grown wafers at Texas State University. The room temperature channel transport data of the antimonide-based pre-MOSFETs is shown in Table 5.4. Wafer 7-133 presents with a hole mobility that is 72% higher compared to the mobility



(a) 7-133 (b) 7-136 (c) 7-137

Figure 5.12: Cross-sections of antimonide-based pre-MOSFETs: (a) lattice-matched channel, non-inverted design (7-133), (b) strained channel, non-inverted design (7-136) and (c) strained channel, inverted design (7-137).

reported for a lattice-matched, 10 nm thick GaSb QW channel [75]. Good confinement of holes provided by the valence band offset of 0.43 eV is reflected by the channel concentration of $3{\times}10^{12}\,\mathrm{cm}^{-2}$ obtained in wafer 7-133. The introduction of 1.21% strain in the channel, in wafer 7-136, results in a 49% improvement in mobility over the lattice-matched pre-MOSFET and further comparable to the best reported mobility values realised in GaSb and InGaSb QW channels subject to 0.5-1.25% strain [74, 177]. The channel density, however, is observed to be reduced by 48% in wafer 7-136 compared to wafer 7-133. A similar finding of decreasing channel density with increasing strain was also noted in [74]. The reduced charge transfer is attributed to the increase in the energy of the valence band states in the QW with increasing strain in the channel [74]. Under an identical strain of 1.21% the inverted growth, wafer 7-137, shows a lower mobility, reduced by 25% from wafer 7-136. This is partly due to dopant segregation, typical of inverted structures, as

Table 5.4: Room temperature Hall data of antimonide-based pre-MOSFETs.

Material	μ_h ($\mathrm{cm}^2/\mathrm{Vs}$)	p_s (cm^{-2})	R_{sh} (Ω/sq)
7-133	738	3.0×10^{12}	2738
7-136	1096	1.57×10^{12}	3636
7-137	818	2.4×10^{12}	3144

discussed in Section 5.4.1.3. The other cause of mobility degradation is the thinner spacer employed in wafer 7-137, which in turn enhances ionised impurity scattering of channel carriers. On the other hand, the increased charge transfer arising from the thinner spacer, has resulted in a 52% increase in channel density of wafer 7-137 in relation to wafer 7-136. The transport properties can be further improved in the inverted growth by increasing the strain in the channel.

It is instructive to compare the antimonide-based pre-MOSFET results of this work to that reported in literature. Shown in Figure 5.13 is a plot of room temperature hole mobility as a function of channel sheet density for strained InGaSb [76, 204, 220, 221], InSb [77] and GaSb [74] QWs. The data points corresponding to wafers 7-133, 7-136 and 7-137 are indicated as diamonds. Using the sheet resistance contour lines as a reference, it is observed that the results of this work are competitive to that achieved in literature, with only strained GaSb QW channels showing better transport performance mainly from increased channel density. The enhanced charge transfer is attributed to the confinement from an AlAsSb barrier, which provides for larger band offsets with increasing strain, and to the employment of double-sided doping [74].

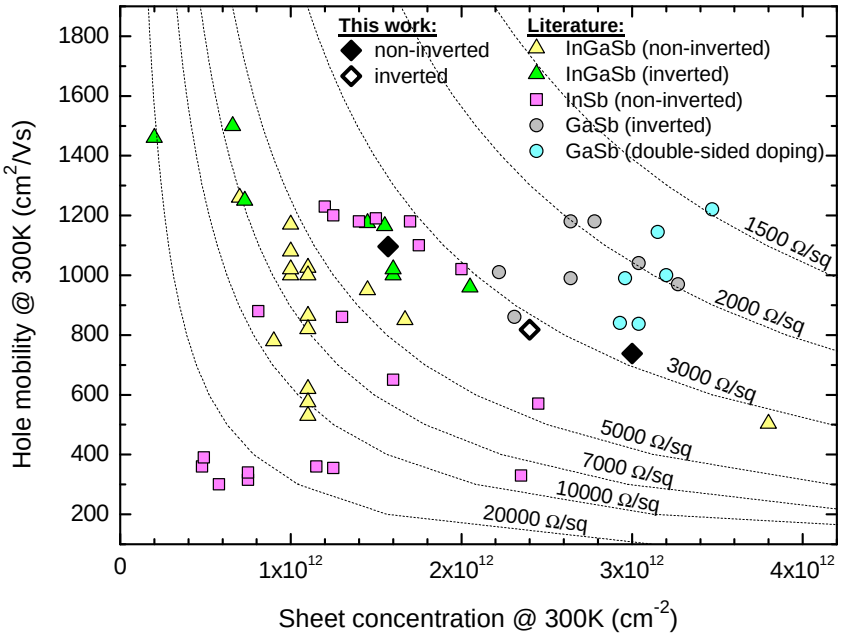


Figure 5.13: Room-temperature hole mobility as a function of sheet concentration for inverted and non-inverted InGaSb quantum wells investigated in this work. Data from literature for strained InGaSb [76, 204, 220, 221], InSb [77] and GaSb [74] quantum wells are also included for comparison.

5.5 Chapter summary

Channel transport, for drive current maximisation, and electrostatic gate control were identified as two aspects of device performance that needed to be addressed in the MOSFET layer design, at the start of the chapter. Strain engineering was also highlighted to be an important technological approach for enhancing the hole mobility of III-V materials. On consideration of the various planar device architectures, the flatband-mode architecture was deemed to be suitable for III-V p-channel MOSFET development in this work. The key design parameters impacting QW channel transport were also discussed. Based on an evaluation of the various arsenide-based and antimonide-based channel/barrier material options available within the III-V material family, $In_{x\geq 0.53}Ga_{1-x}As/In_{0.52}Al_{0.48}As$ and $In_{x\geq 0.1}Ga_{1-x}Sb/AlSb$ were found to be suitable for engineering the QW heterostructure. Following this, the design of the MOSFET layer structure based around these material options was presented. The design parameters investigated for transport optimisation were the spacer thickness, the channel thickness, the channel strain, δ-doping density and doping strategy (inverted or non-inverted). As suitable high-k oxides were not yet available for integration, suitably designed structures, mimicking the electrostatics of III-V MOSFETs, referred to as pre-MOSFETs were used in the investigation. All layers were grown using MBE. Van der Pauw and Hall bar structures were used for room temperature and temperature-dependent Hall characterisation, respectively.

The transport properties associated with arsenide-based pre-MOSFETs were investigated for $In_xGa_{1-x}As$ mole fractions, x, in the range of 53-85% (0-2.1% strain). The lattice-matched, non-inverted structures, regardless of doping level or spacer thickness, exhibited modest room temperature mobilities in the range 60-70 cm^2/Vs. From temperature-dependent Hall analysis, parallel conduction was identified as the cause of mobility degradation. Strained pre-MOSFETs demonstrated excellent transport properties with approximately 4.5-7× hole mobility improvement over the lattice-matched counterparts. Lower mobilities were obtained for the inverted compared to the non-inverted, strained pre-MOSFETs due to enhanced ionised impurity scattering from dopant segregation in the former. Non-inverted pre-MOSFETs subject to 2.1% strain demonstrated the highest room temperature hole mobility of any InGaAs QW channel reported to date.

In the case of the antimonide-based pre-MOSFETs, apart from the doping strategy and channel strain, all other structural/growth parameters were kept identical between the grown wafers. The lattice-matched, non-inverted pre-MOSFET demonstrated a mobility of 738 cm^2/Vs at a channel density of 3×10^{12} cm^{-2}. The induction of 1.21% strain resulted in a 49% mobility enhancement for the non-inverted pre-MOSFET and at the same time

decreased the channel density by the same percentage margin. This decrease in carrier density is due to the valence band states increasing in energy from the addition of strain. For a similar strain, the effects of dopant segregation and the use of thinner spacer resulted in a lower mobility of $818\,\mathrm{cm}^2/\mathrm{Vs}$ for the inverted pre-MOSFET. However, the channel density increased to $2.4\times10^{12}\,\mathrm{cm}^{-2}$ due to the employment of a thinner spacer which gives rise to enhanced charge transfer.

6

Process Modules

6.1 Introduction

There are two stages to realising high performance III-V MOSFETs. The first is heterostructure engineering for optimisation of the channel transport properties, as discussed in the previous chapter. The second is process engineering, which is key to establishing process flows that do not compromise device performance. This stage involves identifying the key modules that make up a device process flow, then investigating them as isolated entities for development and optimisation prior to integration into a complete flow.

A number of process modules were developed and examined for S/D and gate stack engineering during the course of this work as highlighted in Figure 6.1. The motivation for investigating these modules are summarised as

- **Modules (1) and (2):** As devices are scaled to smaller dimensions, the on-resistance in the intrinsic device is reduced. This necessitates the parasitic source/drain series resistance to be a small fraction of the on-resistance in order to maximise the drain current [113] and the transconductance [222]. Therefore, the contact resistance and access resistance, resistance of the gate-to-source/drain (access) regions, need to be minimised in order to achieve high drive currents.

- **Module (3):** Interface defects, border traps and bulk oxide traps are known to reduce the gate modulation efficiency as well as limit Fermi level movement during transistor operation [21, 223]. The result is a degradation of device characteristics both in the on and off-state, and in extreme cases, the device cannot be switched off. Thus, interface passivation technologies need to be adopted to minimise these traps, thereby enabling good device performance to be achieved.

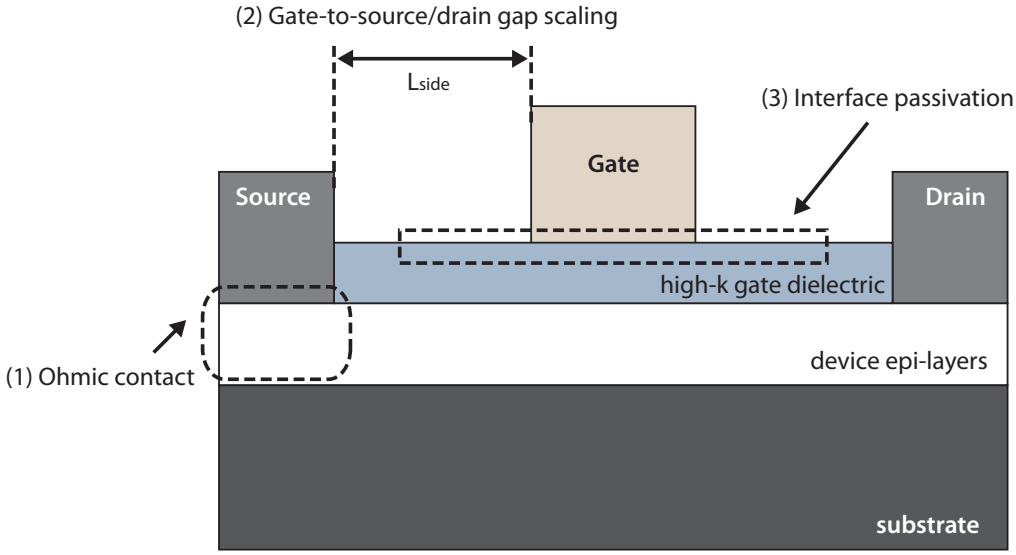


Figure 6.1: Schematic illustration of a flatband III-V MOSFET with process modules identified.

In order to investigate these modules, a baseline device layout, and its accompanying process flow, first need establishing. This then sets the stage for examining the modules as isolated entities. In addition, modifications to the device layout and/or process flow can be addressed as part of module development if the need arises. As such, the next section deals with aspects of baseline device fabrication. The development and optimisation of the isolated process modules are presented in subsequent sections.

6.2 Generic MOS Device Fabrication

In this section the generic processing employed for MOS device fabrication is presented. The description features geometrical layouts as well as process flows adopted for the realisation of MOSFETs and MOS capacitors, taking into consideration process compatibility with the III-V MOS stack and pre-existing substrate features.

6.2.1 MOSFET Fabrication

In this project devices were realised on epi-layer structures based on the flatband architecture that is "normally-on" [94]. This meant that the active (conductive) portion of the fabricated FETs needed to be electrically isolated such that current between the S/D regions is restricted to flow only under the gate (see Figure 6.2(a)). Otherwise parasitic conduction - current flow out with the gated region, not controlled by the gate electrode

- will affect transistor performance detrimentally. In III-V transistor technologies, electrical isolation is predominantly obtained by mesa etching or ion implantation. However such isolation modules were not adopted for transistor fabrication in this work. Instead a "wrap-around" gate topology [182], as illustrated in Figure 6.2(b), was employed which obviates the need for an isolation process. In this device layout, the standard finger design of the gate contact is modified to form a closed ring, encircling the drain contact. As a result the gating action is present along all paths of current flow between drain and source. Additionally, only two levels of lithography, comprising gate and S/D contact definitions, are required to realise a complete device. The result is a rapid turnaround of devices, which is particularly advantageous for a quick assessment of epi-layer design and ideal for MOSFET layer optimisation.

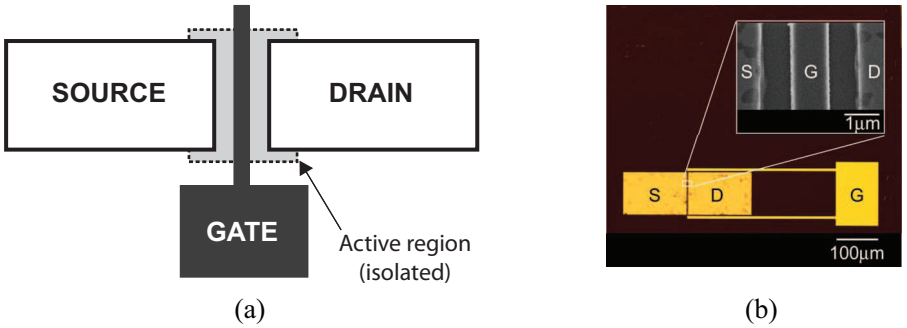


Figure 6.2: Schematic of (a) a FET with an electrically isolated active region and (b) warp-around gate topology [182].

In regards to the processing for "wrap-around" gate devices, a mix-and-match lithographic patterning [123] appears appealing from a cost and throughput perspective. In this method features of critical geometries are patterned with EBL by taking advantage of its superior resolution capability, while optical lithography is used for less critical geometries. In this work transistor gate sizes ranged from hundreds of nanometres to tens of micrometres, while S/D contacts were of the order of 150 μm. This then enables gates and S/D contacts to be patterned by EBL and optical lithography respectively. Pattern overlay accuracy however is limited to $\sim 1\,\mu$m with optical lithography using the Suss MicroTec MA6 mask aligner available in the James Watt Nanofabrication Centre (JWNC). This was deemed unacceptable for alignment between gate and S/D contacts as the devices fabricated in this work have a gate-to-source/drain separation (L_{side}) of $\leq 1\,\mu$m. For these reasons an exclusive EBL process was favoured for device fabrication.

In the fabrication process, markers are written in the same lithographic level as the first pattern layout, to save an additional exposure step as highlighted in Section 3.2.3.4.

Although the chosen S/D contact metals provide excellent contrast to the III-V substrates for marker detection, the thermal processing involved in the gate-last (GL) process may cause roughening of marker edges. This can pose problems for accurate marker detection, thereby resulting in misalignments of pattern overlays. Hence, a gate-first (GF) process flow was exclusively used in this work.

The GF exclusive EBL process flow for "wrap-around" gate device fabrication is shown in Figure 6.3. Following III-V epi growth, the gate dielectric was deposited prior to any processing for device fabrication. For the MOS devices investigated in this work, Al_2O_3 deposited by ALD served as the gate dielectric. The device fabrication comprised of a combined gate/marker level followed by a S/D ohmic level. The latter consisted of a selective wet etch of the gate dielectric in a dilute KOH-based solution [180], followed by contact metallisation and an ohmic anneal.

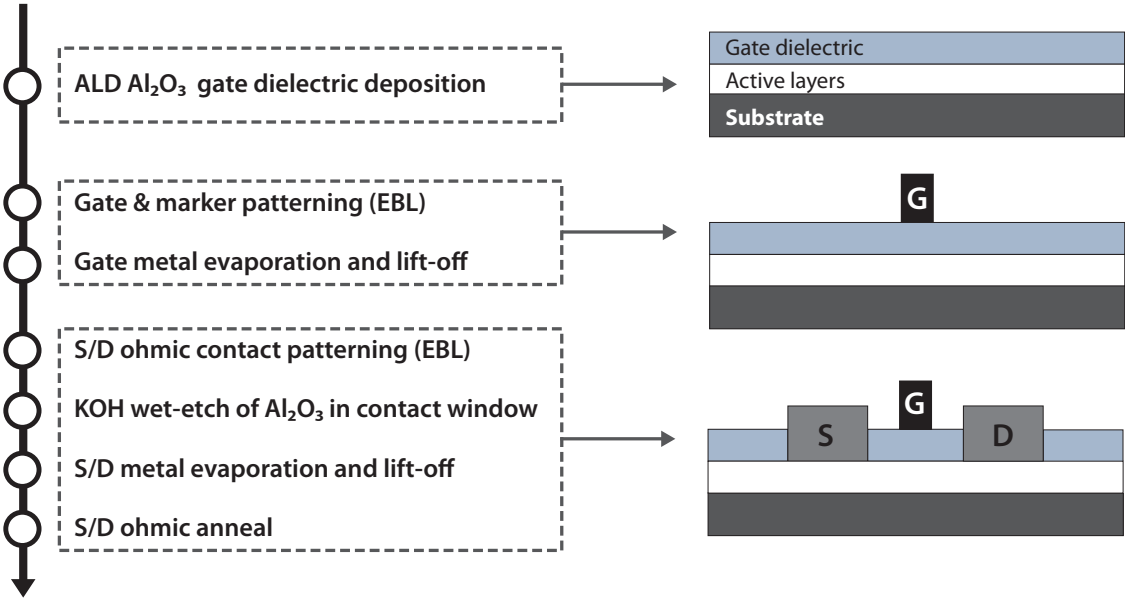


Figure 6.3: Generic process flow for gate-first (GF), lithographically-aligned (LA) MOSFET fabrication employing the "wrap-around" gate device layout, comprising of two levels of EBL.

6.2.2 MOS Capacitor Fabrication

MOS capacitors can be fabricated in one of two ways, depending on the layer structure. Typically MOS capacitors are realised on layers specifically designed for admittance studies. An example of such a structure is illustrated in Figure 6.23, and will be referred to as C–V wafers from hereon in. From bottom-up, it consists of a substrate, epi-layers and

the gate dielectric. The epi-layers are composed of a heavily doped contact layer and a much thicker active layer of lower doping directly beneath the gate dielectric.

Two electrical contacts, gate and ohmic, are required for MOS capacitors in order to characterise the electrical properties of any wafer. For doped substrates a conducting path exists between the back substrate and the active layer. Therefore, a gate contact on top of the dielectric and a metal contact that presents with ohmic behaviour, underneath the doped substrate, would suffice for electrical characterisation as shown in Figure 6.4(a). In terms of the electrode geometries, a circular disk of metal is used for the gate, while a blanket metal layer is used for back ohmic contact. Due to the geometrical requirements the gate metal can be defined either by lithographic patterning or through a shadow mask. Since there is no alignment involved for gate patterning, shadow mask patterning is favourable to lithographic patterning. This simplifies the processing resulting in fewer processing steps, which in turn reduces fabrication time. Additionally, this avoids lithographic processing that could influence the dielectric and its interface with the underlying semiconductor.

Contact configuration on wafers with undoped substrates is slightly more complex. A conduction path between the active layer and the backside of the wafer is non-existent due to the substrate being semi-insulating. Hence, ohmic contacts cannot be realised on the back face of the substrate. Rather, access to the active layer is afforded by forming a conductive path from the top side of the wafer to the edges of the active layer. This is similar to current conduction between the source and drain in a transistor via the channel, which is controlled by the gate. An alternative layout known as the annular C–V structure, as illustrated in Figure 6.4(b), was therefore adopted. In this configuration the ohmic contact is formed as a concentric annular ring, with an extended region for probing purposes, within which a circular gate contact resides. The annular C–V layouts were

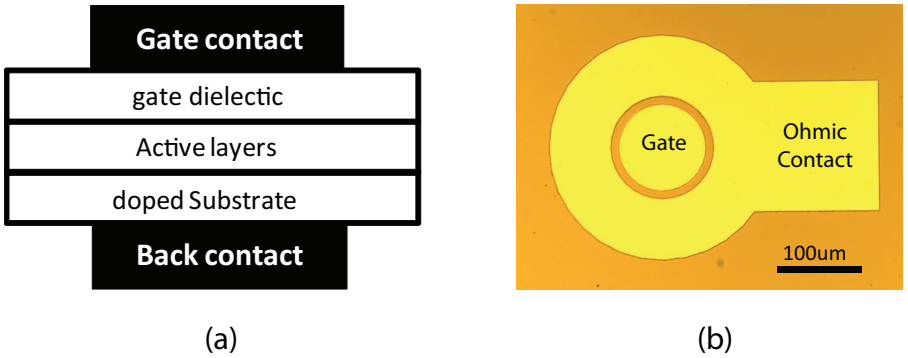


Figure 6.4: Gate and ohmic contact configuration for MOS capacitor realisation on wafers featuring substrates that are (a) doped and (b) semi-insulating.

lithographically patterned for reason of alignment, using the GF process flow used in MOSFET fabrication as shown in Figure 6.3.

In addition to the I–V characteristics, it is important to determine the gate admittance behaviour of transistors fabricated on device wafers. This is for two reasons, the first to enable an assessment of the gate dielectric and the dielectric/III-V interface properties. The other is for mobility extraction (see Section 4.4.2.3). Typically, admittance characteristics of FETs are obtained from split C–V measurements [172]. However, this technique is precluded in the presence of parasitics, which in the case of the wrap-around gate layout arises from the large gate probe pad capacitance [180, 182]. For this reason MOS capacitors based on the annular C–V layout were fabricated alongside MOSFETs on device wafers in this work. The simultaneous processing of the capacitors alongside the transistors also helped to ensure the resulting admittance behaviour was truly representative of the fabricated transistor.

6.3 Ohmic Contacts

6.3.1 Introduction

The realisation of ohmic contacts forms one of the most basic processes in semiconductor device fabrication and is an integral part of any device. The device architecture on which the planar S/D contacts should be formed has a strong bearing on the choice of ohmic contact strategy. While the implant-free nature of the flatband architecture is highly desirable for low thermal budget processing of III-V MOSFETs [224], this also makes achieving low resistance S/D contacts challenging. This is of particular concern in buried channel designs, since the undoped heterolayers above the channel - cap and barrier/spacer layers - add additional resistive elements to contact resistance (R_c), as shown in Figure 6.5.

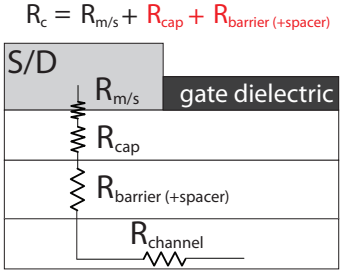


Figure 6.5: Schematic of the S/D region resistor network in the buried channel flatband architecture.

A contact strategy that is commonly employed in heterostructure-based devices is dopant-based metallisation. AuZn metallisation, with Zn as a dopant, is typically used for S/D contacts in p-channel HFETs [187, 193, 197, 202]. Be and Mn also tend to feature as p-type dopants in this metallisation scheme [225, 226]. The high diffusivity of these dopants enables ohmic contacts to be realised on structures with channels as deep as 100 nm from the surface [227]. Based on its suitability for this work, ohmic contacts were developed employing the AuZn contact strategy.

It is worth noting that the aim of the project did not encompass ohmic contact development towards a technological solution, advancement or integration. Rather the work presented in this section was undertaken to develop a contact strategy that would enable the demonstration of III-V flatband p-MOSFETs, thereby providing a means to assess the viability of III-Vs as a p-channel solution in CMOS. It should be recognised that more technologically relevant ohmic solutions would be implemented in III-V CMOS. Regardless, some optimisation of the adopted strategy was attempted to improve R_c, in order to mitigate its impact on device characteristics.

In this section the reaction mechanisms between AuZn metal layers and the underlying semiconductor as well as the technological challenges associated with the contact formation, and mitigating options, are first described. Following this, an overview of the processing techniques for contact realisation, the semiconductor material layers on which the contacts are evaluated and the methodology for assessing the contact performance is presented. The section concludes with the experimental results and discussion.

6.3.2 AuZn Ohmics - An Overview

In order to develop AuZn contacts an understanding of the role played by each constituent element, as well as the chemical and metallurgical reactions with the underlying semiconductor, leading to ohmic behaviour is vital. Barcz *et al.* [228] studied the reaction between Au and GaAs and found the former to act as a catalyst that initiates semiconductor decomposition resulting from arsenic evaporation as follows:

$$\begin{array}{c} Au \\ \downarrow \\ 2\,GaAs \longrightarrow 2\,Ga + As\uparrow . \end{array} \tag{6.1}$$

A reaction of Au and Ga then proceeds as:

$$mAu + nGa \longrightarrow Au_mGa_n. \tag{6.2}$$

This is a self-limiting process; while on one hand the evaporation of As is promoted by the removal of Ga that is in excess, on the other the accumulation of excessive Ga terminates As evaporation [228]. The resulting decomposition of GaAs was reported to occur even at room temperature. In contrast as-deposited Zn does not react with GaAs and requires anneal temperatures of about 320°C to initiate a reaction. In addition, pure Zn shows ohmic behaviour on moderately p-doped GaAs when annealed at temperatures as low as 220°C [225]. However Au on its own, without Zn, is unable to produce ohmic contacts on similar substrates even when annealed at 500°C [225, 228]. This is indicative of Zn, a dopant impurity, being a pre-requisite for ohmic contacts.

AuZn contacts are typically formed from its composite alloy or as a multilayer metallisation, whereby each metal layer is sequentially deposited. Similar to Zn-only contacts, as-deposited AuZn does not react with the underlying semiconductor [229]. While it is to be expected that AuZn composite alloys will be intermixed, such interdiffusion also shows up in as-deposited multilayer metallisation [225, 229]; apart from accumulating at the contact/semiconductor interface, Zn also diffuses into Au to form AuZn phases [225]. For contacts on GaAs substrates, temperatures as low as 100°C are sufficient to cause the following phase transformation [225]:

$$AuZn + Au \longrightarrow Au_3Zn. \tag{6.3}$$

This phase remains stable following annealing at higher temperatures, with no observable reactions at the contact/GaAs interface. However, Zn begins to migrate to the surface of the metal stack, in addition to the aforementioned diffusion behaviour [225]. Although it is recognised that Zn acts as a doping agent, the mechanism by which this occurs is ambiguous since there was no experimental evidence to support the hypothesis of a highly doped GaAs region below the contact [228]. While the creation of Ga vacancies is believed to be a necessary condition of Zn incorporation, significant decomposition of the semiconductor in the formation of the AuGa phase is merely a reaction by-product and therefore not a pre-requisite for realising ohmic contacts [228]. Besides its role as a dopant source, the inclusion of Zn also has other benefits; the decomposition of the semiconductor and outdiffusion of the relevant element is less pronounced and the reaction temperature between Au and the underlying semiconductor is lowered [225].

The reaction temperature for the onset of ohmic behaviour is not only dependent on the inclusion of Zn in the metallisation. Other factors, primarily to do with the semiconductor, have a significant impact as well. This is illustrated in Table 6.1 which compares the temperature at which ohmic behaviour sets in relation to substrate doping, Schottky barrier height (SBH) for holes and effective hole mass for GaAs, InP and GaP. It is immediately obvious that the onset temperatures are greatly different between the semiconductors. While the doping levels are not too dissimilar, the onset temperature for GaP is 120°C higher than in GaAs. This is attributed to the SBH and effective mass being $\sim$2$\times$ and $\sim$1.5$\times$ larger respectively for GaP. The importance of these two parameters is further apparent when considering GaAs and InP. Even an order of magnitude increase in the doping is not able to compensate for the larger barrier height and effective mass of InP. As a result the onset temperature is still higher for InP when compared to GaAs.

Table 6.1: Onset of ohmic behaviour in GaAs, InP and GaAs in relation to substrate doping, hole Schottky barrier height (ϕ_{Bp}) and effective mass of holes (m_{h}). Data for doping, specific contact resistance (ρ_{c}) of contacts annealed at 420°C for 180s and ϕ_{Bp} are adapted from [229]. m_{h} data is from [68].

Semiconductor	Doping (cm^{-3})	ρ_{c} ($\Omega\mathrm{cm}^2$)	ϕ_{Bp} (eV)	m_{h}	Onset of ohmic behaviour (°C)
GaAs	2×10^{17}	5.5×10^{-4}	0.4	$0.51m_{\mathrm{o}}$	350
InP	5.5×10^{18}	1.8×10^{-4}	0.85	$0.6m_{\mathrm{o}}$	380
GaP	6×10^{17}	9.5×10^{-4}	0.9	$0.79m_{\mathrm{o}}$	470

Traditionally AuZn contacts are formed from vacuum evaporation of the composite alloy and subsequent heat treatment to realise ohmic characteristics. However, there are a number of technological problems associated with such deposition and annealing. During evaporation Zn preferentially evaporates before Au due to the difference in vapour pressure between the two elements. This coupled with the Zn sticking coefficient being extremely low leads to adhesion issues as well as non-homogeneity of the resulting metallisation [230]. Other issues with AuZn include outdiffusion/evaporation of semiconductor elements, indiffusion of Au [225, 226, 228] and re-evaporation of Zn [230] during contact annealing. Such effects are detrimental to contact morphology and R_{c}. A couple of measures have been employed during contact preparation to alleviate these problems. Sequential deposition of the individual metals comprising AuZn would compensate for the vapour pressure difference. Nonetheless, adhesion issues would still persist with Zn being

deposited as the first layer. This can be overcome by employing thin metal layers, for example Au, prior to AuZn deposition to serve as nucleation layers for Zn condensation [230].

The problems of outdiffusion, indiffusion and evaporation of materials comprising the semiconductor and ohmic metallisation can be curbed using a number of different techniques, and are discussed in relation to GaAs. The use of As overpressure during annealing can restrain the evaporation of As. RTA as opposed to furnace annealing has been suggested to mitigate these issues [231]. Annealing the contact with dielectric-capping layers, such as SiO_2, Si_3N_4 and Al_2O_3, prevents the sublimation of As and Zn [226]. Since Au is in fact a catalyst in the self-limiting reaction leading to GaAs decomposition, thinner metallisations would act to reduce As evaporation, Ga outdiffuison as well as excessive Au indiffusion [226]. Another technique involves the insertion of diffusion barriers in between the first layer of AuZn alloy and an outer Au layer of much greater thickness. Metal such as Ni and Ti have been found to be effective as diffusion barriers in mitigating the undesirable effects of the ohmic-semiconductor reaction [230, 232].

6.3.3 Experimental Details

AuZn contacts were developed using pre-MOSFET wafers as MOSFET material was not available at the time of investigation. TLM structures (see Figure 4.4(a)) were fabricated for contact performance evaluation. Device processing comprised patterning of TLM contact pads, using either optical lithography or EBL, followed by an additive pattern transfer process comprising metal deposition and lift-off (see Section 3.3.2). Prior to loading into the thermal evaporator, the patterned wafers were treated in a dilute HCl solution (1 $HCl : 4\,H_2O$, 30 s) for native oxide removal and rinsed for 30 s in de-ionised water, unless otherwise stated. Contact metallisation was prepared by sequentially evaporating the individual metal layers from resistance-heated crucibles. The pressure during evaporation was in the range of $1\text{–}4\times10^{-6}$ mbar. Following ohmic metal lift-off, the alloying process was performed in a strip heater with a temperature precision of ±1 °C under continuous flow of forming gas (5% H_2/ 95% N_2).

For R_c evaluation, linear regression was performed on resistance data obtained as a function of gap spacings. TLMs with a minimum of four gap spacings were used for data fitting, and the measure of the goodness-of-fit between the data and the gap sizes of a single TLM structure was determined by the correlation coefficient. The correlation value has to be as high as possible, preferably $\geq$0.99. This criteria is used to determine if a TLM dataset is valid and useable. In order to assess the homogeneity of contact performance,

R_c was compared and averaged over a minimum of four TLM sets electrically characterised at different positions on a sample.

6.3.4 Results and Discussion

Ohmic contacts were developed using a trilayer Au/Zn/Au metallisation. Based on the review presented in Section 6.3.2, a 5 nm Au nucleation layer was adopted as the first metal layer to be deposited. A series of experiments were formulated to investigate the effect of the Au/Zn/Au metallisation thickness, incorporating a Ti layer as a diffusion barrier within the contact metallisation and de-oxidation conditions on the resulting contact performance.

6.3.4.1 Effect of Au(5nm)/Zn(Xnm)/Au(Ynm)

The initial investigation of Au/Zn/Au ohmics began with the experimentation of the Zn and the upper Au layer thickness using c674 pre-MOSFET wafer illustrated in Figure 6.6. The dependence of R_c on Au(5 nm)/Zn(X)/Au(Y) thickness and anneal time at a fixed temperature is shown in Table 6.2. It was found that increasing the thickness of the Zn and upper Au layer at a fixed annealing time of 60 s improved R_c by ∼28%. This improvement is likely due to the increased Zn incorporation from the metallisation with the thicker Zn metal layer. Moreover, the thicker Au layer could have led to increased Ga vacancies as highlighted in Section 6.3.2, thereby enhancing Zn incorporation into the underlying semiconductor.

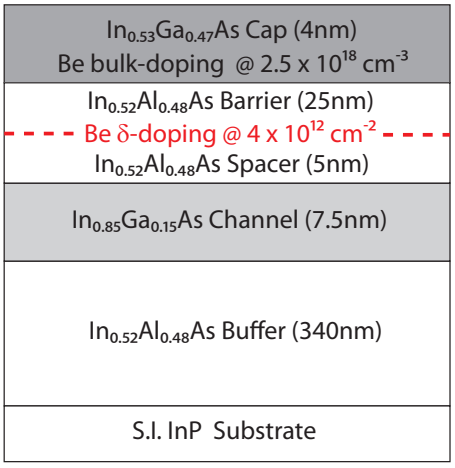


Figure 6.6: Cross-section of arsenide-based pre-MOSFET, c674.

As evident from Table 6.2 the metallisation thickness as opposed to anneal time has a larger bearing on contact performance. Nonetheless, contact resistance values were still high, with the best performing metallisation of Au(5 nm)/Zn(35 nm)/Au(125 nm) annealed for 120 s, presenting with R_c of 13.33±0.99 Ω.mm. For this reason the observed trend of better contact performance from increasing metallisation thickness was investigated further.

Table 6.2: R_C of Au/Zn/Au contacts on c674 pre-MOSFET, annealed at 400°C.

Contact metallisation	t_c (sec)	R_c (Ω.mm)	Correlation
Au(5 nm)/Zn(20 nm)/Au(70 nm)	30	20.44 ± 0.62	0.9980 ± 7.6 $\times 10^{-4}$
	60	21.52 ± 1.04	0.9983 ± 4.3 $\times 10^{-4}$
Au(5 nm)/Zn(35 nm)/Au(125 nm)	60	15.55 ± 1.54	0.9956 ± 2.4 $\times 10^{-3}$
	120	13.33 ± 0.99	0.9976 ± 6.9 $\times 10^{-4}$

Figure 6.7 depicts the dependence of R_c for a fixed Au(5 nm)/Zn(40 nm)/Au(150 nm) thickness as a function of anneal time and temperature on c674 pre-MOSFET. For the 120 s anneal, the lowest R_c of 12.992±0.43 Ω.mm was obtained at 380°C, with lower and higher temperatures showing slightly degraded contact performance of ∼13.5 Ω.mm. These values are not too dissimilar from Au(5 nm)/Zn(35 nm)/Au(125 nm) metallisation annealed for 120 s at 400°C (see Table 6.2). For the same metallisation thickness, a longer annealing time of 180 s resulted in degraded contact performance across the investigated temperature range in comparison to the 120 s anneal. The poorer performance could be attributed to an increased alloying effect between the contact metals and the semiconductor layers. It is plausible that such an effect causes enhanced decomposition of the underlying semiconductor, leading to increased As evaporation and formation of AuGa phase, which are detrimental to contact performance. In addition the migration of Zn to the surface of the metal contact upon annealing, as highlighted in Section 6.3.2, could have been enhanced by the longer anneal time, thereby causing less Zn to be incorporated into the semiconductor. As such the higher R_c values could be a consequence of the reduced doping in the semiconductor.

A thick upper Au layer could cause an increased decomposition of the underlying semiconductor leading to an increase in As evaporation, which acts to degrade the performance of the contacts [226]. Therefore, it was deemed necessary to investigate a thinner upper Au layer and its associated effect on R_c. Such a study was undertaken on c665 pre-MOSFET (see Figure 6.8) for upper Au layer thicknesses of 100 nm

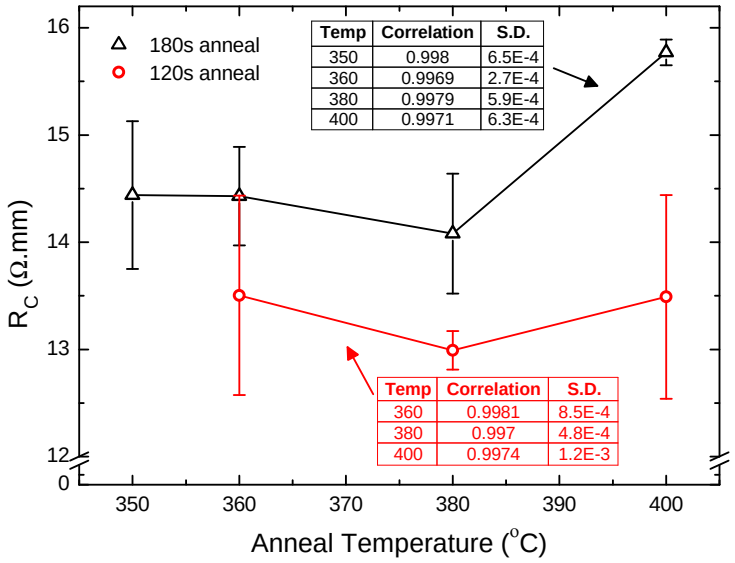


Figure 6.7: R_c comparison of Au(5 nm)/Zn(40 nm)/Au(150 nm) contacts on c674 pre-MOSFET, annealed at 350-400°C for 120 s and 180 s.

and 150 nm, the results of which are reported in Figure 6.9. While the metallisation with the 150 nm thick upper Au layer showed comparable contact performance to that obtained on c674 pre-MOSFET under identical process conditions, the contact performance of metallisation comprising 100 nm thick upper Au layer was severely degraded. At 380°C, R_c of Au(5 nm)/Zn(40 nm)/Au(150 nm) was 13.498±0.43 Ω.mm, while that of Au(5 nm)/Zn(40 nm)/ Au(100 nm) was 22.747±0.76 Ω.mm, translating to a 70% increase in contact resistance. This then suggests that the 100 nm thick upper Au layer is not adequate in creating sufficient Ga vacancies for Zn incorporation.

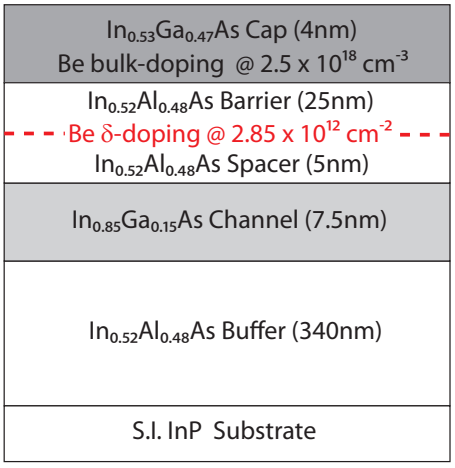


Figure 6.8: Cross-section of arsenide-based pre-MOSFET, c665.

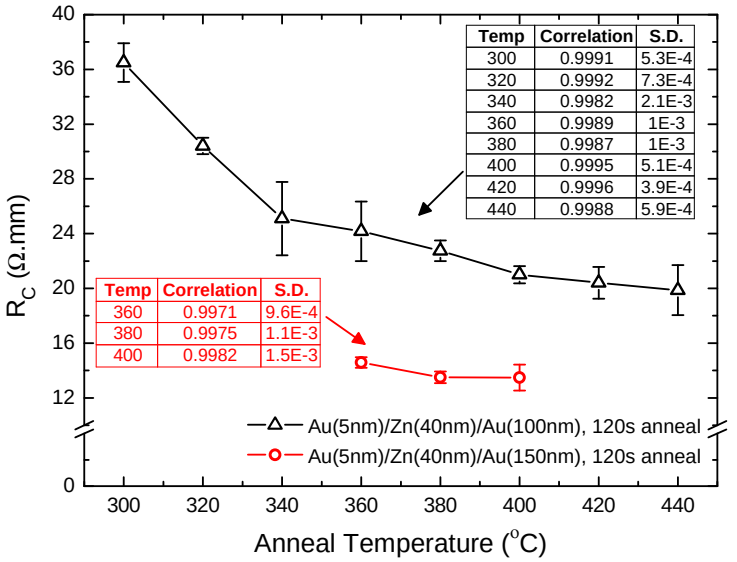


Figure 6.9: R_c comparison of Au/Zn/Au contacts on c665 pre-MOSFET for various upper Au layer thicknesses and annealing temperatures, for a fixed annealing time of 120 s.

6.3.4.2 Effect of Diffusion Barrier

In an effort to optimise the performance of the contacts developed in this work, Ti metal was investigated as a diffusion barrier. This was motivated by the work in [232] in which Ti was reported as an effective barrier against excessive Au diffusion to the semiconductor, a know mechanism for contact resistance degradation. The suppression of Au migration stems from Ti diffusing more quickly to the contact/semiconductor interface during annealing. It was further suggested that Ti aids in creating Ga vacancies at the interface, thereby facilitating Zn incorporation, which would also explain the improvement in R_c [232].

The efficacy of Ti as a diffusion barrier was investigated via e-beam evaporation of Ti/Au following on from the thermal evaporation of Au/Zn/Au on c674 pre-MOSFET. The results of the contact performance with and without a diffusion barrier are illustrated in Table 6.3. One noteworthy fact is that R_c of metallisation without the diffusion barrier showed improvement over the previously reported value of 12.992 Ω.mm (see Figure 6.7) even though all conditions, including the pre-MOSFET material, were identical. This highlights the issue of run-to-run variations for the Au/Zn/Au metallisation process. A possible explanation is the variability of the deposition rate during Zn evaporation witnessed for all metallisation runs. The expected enhancement in the contact performance was not observed with the inclusion of a diffusion barrier. Hirano *et al.* [232] observed

Table 6.3: R_C comparison between Au/Zn/Au and Au/Zn/Au/Ti/Au contacts on c674 pre-MOSFET, annealed at 380°C for 120 s.

Contact metallisation	R_c (Ω.mm)	Correlation
Au(5 nm)/Zn(40 nm)/Au(150 nm)	10.77 ± 0.46	0.9997 ± 1.3 $\times 10^{-4}$
Au(5 nm)/Zn(40 nm)/Au(150 nm)/Ti(100 nm)/Au(100 nm)	11.36 ± 0.49	0.9996 ± 1.8 $\times 10^{-4}$

that the thickness of Ti was an important consideration towards contact performance enhancement, and reported degraded R_c values for too thin a Ti layer and for an excess of Ti. This was attributed to the thinner layers being inadequate in suppressing Au diffusion, while the thicker layers began to suppress Zn incorporation into the underlying semiconductor, a necessary condition for good ohmic behaviour [232]. This is suggestive of the Ti thickness used in this study (100 nm) being non-optimal. Additionally, the break in vacuum between Au/Zn/Au and Ti/Au depositions, as a consequence of the exchange of wafers between the evaporators, could have contributed to the degradation of R_c.

6.3.4.3 Effect of De-Oxidation Conditions

It was important to assess the performance of Au/Zn/Au contacts on antimonides since devices were also realised using these materials. One particular issue is native oxides that tend to grow to several nm in thickness on antimony surfaces, as highlighted in Section 6.5.1. Thus, the surface preparation condition for native oxide removal prior to metal deposition, a step known as de-oxidation, could have a strong influence on the contact performance. For this reason two different de-oxidation conditions were examined on 7-137, an antimonide-based pre-MOSFET wafer (see Figure 6.10). The first was a

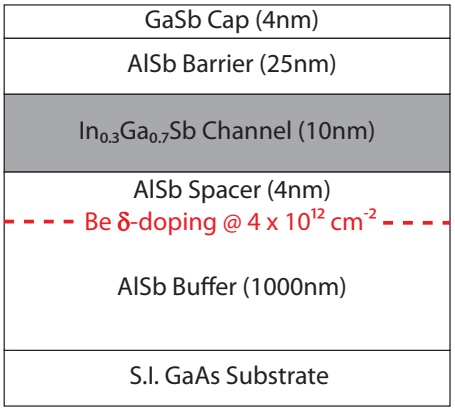


Figure 6.10: Cross-section of antimonide-based pre-MOSFET, 7-137.

30 s dip in 1 HCl : 4 H_2O solution, a de-oxidation solution typically adopted for InGaAs surfaces. This will be referred to as de-ox1 hereafter. A less diluted solution (1 HCl : 2 H_2O) and a longer dip time of 60 s comprised the second condition (dex-ox2).

The resulting contact performance, as shown in Figure 6.11, was noticeably different between the two de-oxidation conditions. In the case of de-ox1, although R_c (4.808±5.35 Ω.mm) showed significant improvement over contacts realised on arsenide-based pre-MOSFETs discussed in Section 6.3.4.1, the standard deviation (S.D.) was considerably large. In addition the correlation coefficient was well below the value of 0.99 required of a reliable measurement. On the other hand, metallisations that were subjected to de-ox2 showed much improved, reliable contact performances, with correlation coefficients >0.99. The lowest contact resistance of 1.862±0.35 Ω.mm was obtained for an anneal temperature of 340 °C. The drastic improvement in the contact performance implies that de-ox2 is better suited for native oxide removal on GaSb. It is further noted that R_c of metallisations on this material system shows at least an order of magnitude improvement over that of InAlAs/InGaAs. Such a drastic improvement is to be expected given that SBH is approximately 4–5× smaller for GaSb in comparison to InGaAs. The smaller GaSb effective mass of holes is an additional advantage.

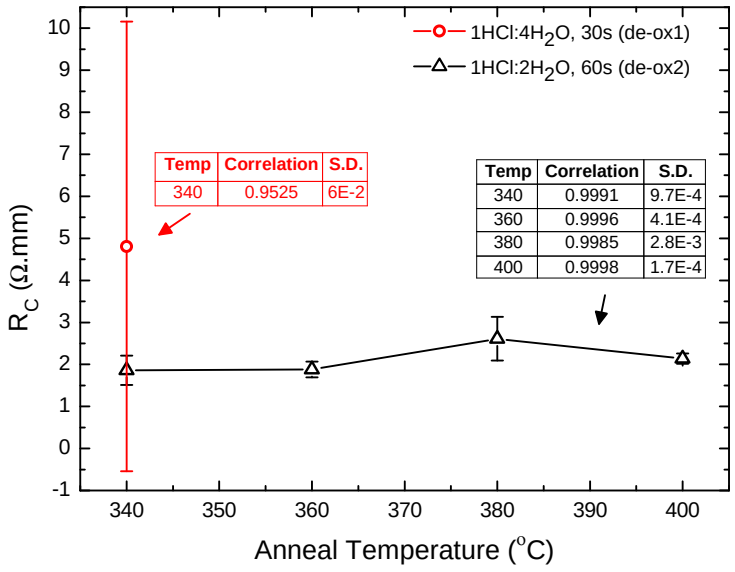


Figure 6.11: R_c comparison of Au(5 nm)/Zn(40 nm)/Au(100 nm) contacts on 7-137 pre-MOSFET annealed at 340-400°C for 120 s under different de-oxidation conditions.

6.3.5 Summary

The development and optimisation of ohmic contacts was challenging due to the implant-free nature of the device architecture. Based on its suitability, a Au/Zn/Au ohmic contact strategy was employed for investigation on pre-MOSFETs with InGaAs and GaSb cap layers. The investigation encompassed the effects of metallisation thickness, a diffusion barrier and de-oxidation conditions. The best performing R_c of 10.77±0.46 Ω.mm on InGaAs pre-MOSFET was obtained for Au(5 nm)/Zn(40 nm)/ Au(150 nm) metallisation annealed at 380°C for 120 s, under a de-oxidation condition of 1 HCl : 4 H_2O, 30 s prior to metallisation.

Contacts to GaSb pre-MOSFETs were found to be more sensitive to de-oxidation conditions. In contrast to InGaAs, a more concentrated solution (1 HCl : 2 H_2O) together with a longer dip time (60 s) was found to be more effective for de-oxidising the surface of GaSb. Contacts, similar to that implemented on InGaAs, performed significantly better on GaSb pre-MOSFETs, with a low R_c of 1.862±0.35 Ω.mm realised for Au(5 nm)/Zn(40 nm)/Au(100 nm) metallisation annealed at 340°C for 120s. Reasons for the difference in contact performance between the two material systems were highlighted.

Attention was also drawn to the issue of run-to-run variability of Au/Zn/Au contact performance under identical processing conditions, which was found to have implications on the device design methodology in the next section.

6.4 Lateral Device Footprint Scaling

6.4.1 Introduction

Parasitic series resistance associated with S/D regions, R_{sd}, can have a sizeable contribution to the overall device resistance, resulting in performance degradation as discussed in Section 2.5. It is therefore apparent that S/D engineering, for the alleviation of R_{sd}, is an essential step towards harnessing the full potential of MOSFETs. Key to this is an appreciation of the individual parasitic components comprised within the source and drain regions, as identified in Figure 6.12.

The total parasitic S/D resistance of a device is given by:

$$\begin{aligned} R_{sd} &= 2R_c + 2R_{side} \\ &= 2R_c + 2\left(\frac{L_{side}}{W} R_{sh}\right), \end{aligned} \tag{6.4}$$

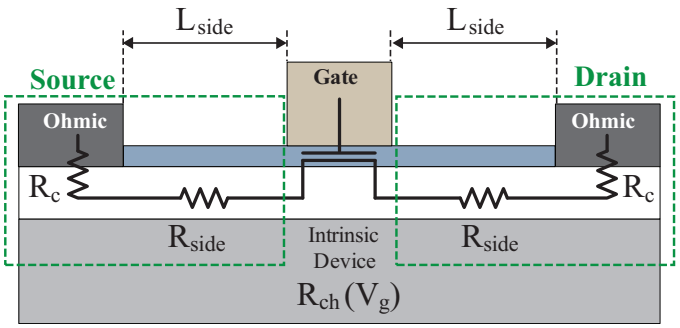


Figure 6.12: Cross-section of a generic MOSFET illustrating the channel resistance (R_{ch}) and the individual components - contact resistance (R_{c}) and access resistance (R_{side}) of the gate-to-source/drain gaps - that make up the source and drain resistances.

where R_{c} is the contact resistance, access resistance (R_{side}) is the gate-to-source/drain resistance (access resistance), L_{side} is the gate-to-source/drain separation, W is the normalised width and R_{sh} is the sheet resistance of the material in Ω/sq. Based on Equation (6.4), series resistance can be minimised through the reduction of the contact resistance, as dealt with in the preceding section, and/or access resistance. The latter resistance has a geometrical and material dependence. Therefore, by scaling down the geometrical parameter, L_{side}, the access resistance contribution can be drastically reduced. As the baseline devices fabricated in this work featured a 1 μm L_{side}, there was scope to downscale the gap spacing into the sub-100 nm regime.

In view of the benefit to the key device metrics of $I_{\mathrm{d,sat}}$ and g_{m}, deep submicron L_{side} scaling was investigated in this work. The investigation was broken down into three phases. In the first phase, processing techniques appropriate for the intended task were considered, following which a method was chosen for implementation. The chosen method underwent development in the second phase to arrive at the requisite conditions required to achieve highly scaled gaps. The investigation was concluded with the final phase, which was the qualification of the method on device material. These different stages of the investigation are discussed next.

6.4.2 Design Phase

6.4.2.1 Self-Aligned vs. Lithographically-Aligned Techniques

Both self-aligned (SA) and lithographically-aligned (LA) fabrication techniques were considered for submicron gap (L_{side}) scaling between the gate and S/D contacts. The suitability of these techniques were assessed based on their compatibility with the processing criteria of the p-MOSFETs developed in this work: aluminium (Al), with an effective

work function (EWF) of 3.9 eV [233], as gate contact to achieve the desired threshold voltage for E-mode device operation; alloyed Au/Zn/Au ohmic contacts; minimal physical/chemical damage to both the gate dielectric and III-V epi-layers from deposition and patterning processes [132, 143, 234–237].

Self-aligned Techniques

SA techniques are commonplace in Si CMOS. The gate and sidewall spacers combined serve as a mask for implantation and silicide formation during frontend-of-the-line (FEOL) and backend-of-the-line (BEOL) processing respectively. This leads to the inherent self-alignment of the S/D implants and silicide, the latter referred to as salicide (SA silicide) [48].

SA techniques, somewhat analogous to that adopted in CMOS, are also employed in III-V technology. HEMTs in particular make use of the T-gate head as a masking layer for S/D contact formation. The T-gate profile causes a discontinuity of the ohmic metal, blanket deposited across the entire device, and in the process enables S/D contacts to be defined in a SA manner as shown in Figure 6.13. While a L_{side} of 125 nm has been realised via this method [238], further scaling has been hampered by the size of the gate-head which dictates the degree of self-alignment.

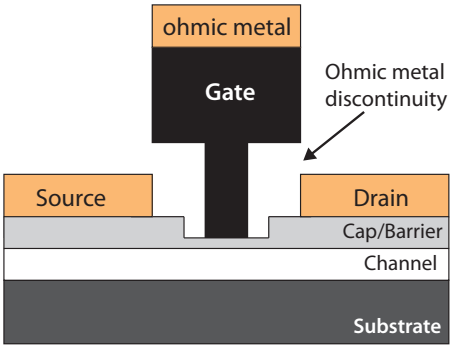


Figure 6.13: Conventional self-aligned III-V HEMT.

Alternatively, aggressive L_{side} scaling to 30 nm has been reported for an "air spacer" process [239] for which self-alignment is attained independent of the size of the gate-head. In this method, a dielectric/ohmic bilayer is etched to define the gate foot, following which the ohmic metal is pulled back through the process of selective isotropic plasma etching, prior to gate metal deposition. As a result, L_{side} is determined by the distance of metal pull back from the edge of the gate foot, which can be controlled through the etch process. While this appears to be an elegant solution, non-alloyed, single metal ohmic contact is a central requirement to the process. This ensures surface planarity and provides the edge

definition needed for aggressive L_{side} scaling [239]. However, the requirement of alloyed Au/Zn/Au ohmics for devices fabricated in this work renders this SA solution unfeasible.

A fully SA scheme that is compatible with the aforementioned p-MOSFET criteria was previously developed at the University of Glasgow for III-V n-MOSFET fabrication [240]. In this scheme, that combines techniques from III-V and Si processing, a Si_3N_4 sidewall spacer process is used to define L_{side}. By tuning the thickness of the conformal layer of Si_3N_4 for sidewall spacer definition the spacer thickness, and therefore L_{side}, can be adjusted to specification; 70 nm thick spacers have been realised from a 300 nm Si_3N_4 film [240]. Low-damage subtractive patterning via plasma etching is a highlight of the scheme and is used for both gate and spacer definition. Additionally Au/Zn/Au is compatible with both dry and wet etching [241] for the removal of adjoining metal between S/D contacts during SA ohmic deposition, which is a necessity to prevent device short circuiting. The only drawback is the spacer material; based on preliminary screening, Si_3N_4 as an encapsulation layer was found to degrade the transport properties of p-MOSFETs developed in this project. As a result the device performance benefits expected from scaling L_{side} would be compromised by using Si_3N_4 as a spacer technology. As a suitable spacer material had not been identified at the time of the scaled L_{side} process development, this scheme could not be utilised.

Lithographically-aligned Techniques

Since none of the SA schemes seemed viable for L_{side} scaling, techniques based on lithographic alignment were next explored. Source-drain gaps of 1 μm or larger have been routinely realised using EBL in an additive pattern transfer process of metal deposition and lift-off [180, 182]. However, the realisation of two large S/D contact pads (150 μm$\times$100 μm) separated by a deep submicron gap, that is uniform along the pad width, is a challenging prospect using EBL. Previous attempts at such a fabrication process, at the University of Glasgow, encountered difficulties. It was found the gaps suffered from resist thinning from proximity effect, with thickness loss worsening for decreasing gap size; the resist was thinned by 80 nm (1.6 μm gap) to as much as 200 nm (270 nm gap). Such resist loss, especially for the smallest gaps, resulted in unreliable ohmic metal lift-off [242].

As discussed in Section 3.2.3.2, electron back scattering arising from the exposure of adjacent patterns will significantly add to exposure of both the patterned and unpatterned resist regions. On GaAs and InP substrates almost half of the exposure dose of large area patterns comes from back scattered electrons [243]. Since the gap resist is situated between two such large area patterns, it is also exposed to at least 50% of the clearing dose [242]. This results in a decrease of the resist contrast which in turn

causes the narrowing, or worse still the closure, of such small gaps. Correction softwares typically applied to compensate for the proximity effect has little value in this case. This is because the exposure dose around the gap regions is uniform and remains unchanged even after proximity correction is applied [244]. Methods for improved resist contrast such as cold development [127], and developer mixture comprising IPA and water as a developer [128], were also found to have little impact on the fabrication process reported in [242]. Ultimately, small gaps were successfully realised by defining the metal pads in separate lithographic steps [242].

The writing of contact pads in separate lithographic levels is one of three approaches investigated by Thoms *et al.* [244] in fabricating small gaps between two large area patterns, all based on the additive pattern transfer technique of metal deposition and lift-off. This double patterning approach using EBL completely avoids the issue of proximity effect and therefore has the potential for ultimate gap scaling. Instead, focus is centred around accurate alignment. The second technique makes use of a HSQ/PMMA bilayer resist. As opposed to pad writing in double patterning, the gap itself is written in HSQ, which is a high-resolution negative tone e-beam resist capable of sub-10 nm patterning [122]. Prior to pattern transfer by metal deposition and lift-off, PMMA is removed in HSQ-free regions by dry-etching. Pattern transfer however leaves the entire sample covered in metal, apart from the written areas. Therefore, to avoid device shorting, the borders around the ohmic pads must be isolated. In the third and final method, a trilayer stack comprising a high contrast positive tone UVIII resist, Si_3N_4 dielectric and a sacrificial layer of polydimethylglutarimide (PMGI) (UVIII/Si_3N_4/PMGI) was used. Following exposure and development of UVIII for contact pads, the nitride layer and PMGI were dry-etched for subsequent pattern transfer.

In comparing all three approaches, the double patterning method stood out as being most viable for adoption. From a processing perspective the other two approaches appeared somewhat more involved, whereas double patterning would merely be an extension to the already established baseline process flow used in device fabrication (see Figure 6.3). Moreover, dry-etch processing in the course of ohmic patterning is not ideal and thus renders the other approaches unfavourable.

6.4.2.2 Design Methodology

One caveat with the aforementioned double patterning approach is the possibility of differing ohmic performance between the source and drain contacts, arising from the separate metallisation steps. It was highlighted in Section 6.3.4.2 that the contact performance of

the Au/Zn/Au metallisation investigated in this work varies from run-to-run. Therefore, it was essential for S/D contact metallisations to be defined concurrently.

One way of circumventing the issue surrounding concurrent pad definition is to limit the lithography to well within the electron back scatter range. This ensures the back scatter contribution to resist exposure is relatively even in all regions. Therefore, for concurrent S/D contact definition, the geometrical pad length needed to be smaller than the back scatter range, with the width fixed at 100 μm. At the maximum VB6 exposure voltage of 100 kV, the back scatter range of the substrates used for device layers in this work is very similar; 12.65 μm for InP and 12.78 μm for GaAs [243]. Within these back scatter ranges, it is further advantageous to use as small a pad length as possible. Two factors were taken into account in determining the smallest pad length that can be adopted in the process, these are discussed in the following paragraphs.

In order to perform measurements a significantly larger S/D pad needed to be patterned and aligned to overlap the geometrically smaller S/D contact. Hereon in, the aforementioned smaller and larger S/D contact pads will be referred to as stripe and large contacts respectively. In the current baseline process flow that employs concurrent patterning of 150 μm$\times$100 μm S/D contacts, L_{side} gaps $\geq$ 1 μm are easily achieved. Therefore, the smallest distance from the gate edge at which point an overlap can be made is 1 μm. Taking this into account, stripe contacts need to be greater than 1 μm in pad length to achieve L_{side} gaps in the tens of nm range.

When down scaling the stripe contact length, it is also important to ensure the ohmic contact performance is not compromised. As highlighted in Section 4.3, the transfer length of contacts defines the contact resistance. Transfer length represents the distance from the contact edge over which 1/e of the total current flows from the substrate to the metal. For efficient ohmics, the physical contact length needs to be at least twice as long as the transfer length. The transfer length of Au/Zn/Au ohmics evaluated on InGaAs and GaSb pre-MOSFETs were estimated to be 1.03 μm and 0.47 μm respectively. This restricts the smallest stripe contact length to about 2 μm, which in essence should behave similar to a longer contact. Nonetheless, this was not a crucial consideration. The reason being, the inclusion of the large contacts as part of the double patterned S/D ohmics ensures the resulting ohmic performance is representative of long contacts.

Based on the preceding arguments, stripe contacts with a pad length of 2 μm was decided upon. Large contacts would then be aligned to the stripe contacts, such that the two contacts are overlapped over a distance of 1 μm. The device layout featuring the double patterned S/D contact design is illustrated in Figure 6.14. It can be seen that the

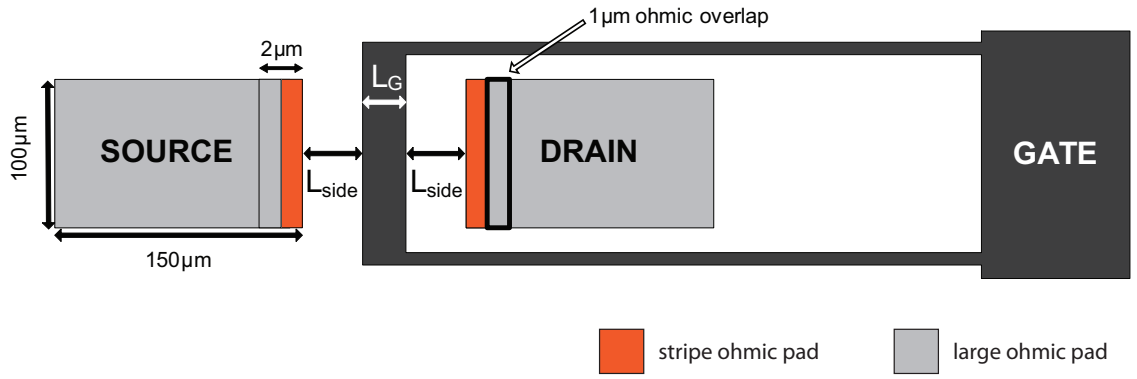


Figure 6.14: Plan view of modified "wrap-around" gate device layout featuring double patterned S/D contact design (not to scale).

new layout is essentially a modification of the existing "wrap-around" gate device layout.

The process flow for fabricating scaled L_{side} devices utilising the modified device layout is shown in Figure 6.15. To accommodate patterning of stripe contacts the baseline device flow is modified to include an additional lithographic level resulting in three levels of EBL. Stripe and large contacts are fabricated using similar processing steps of EBL, wet etch

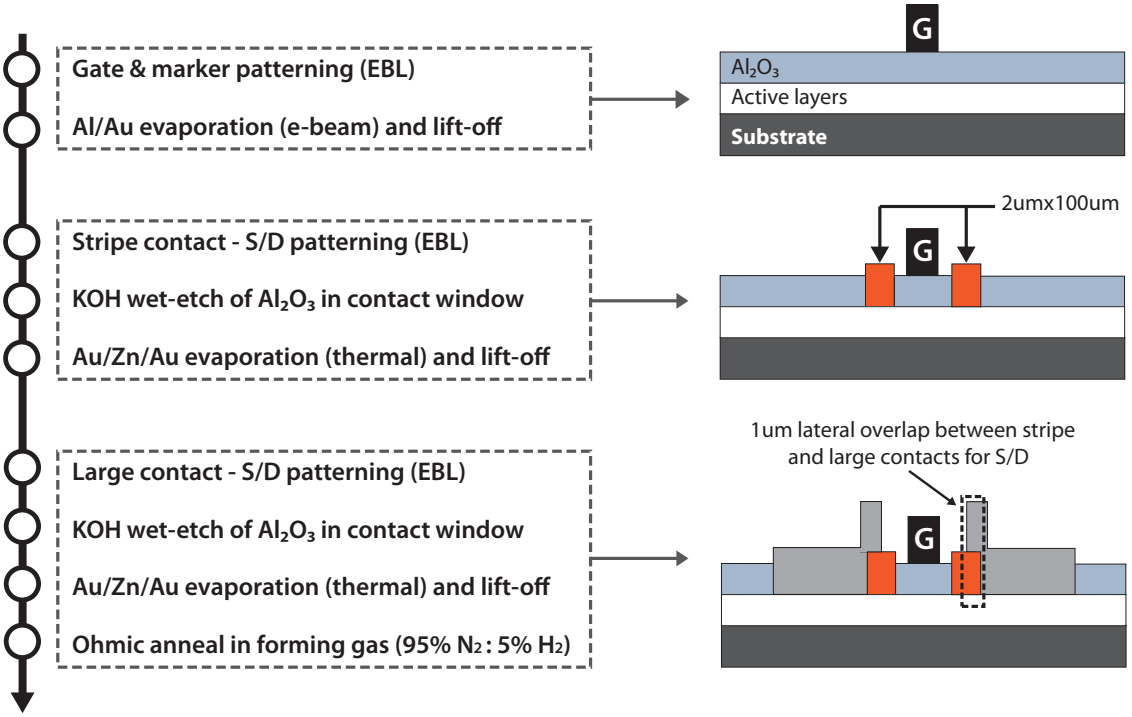


Figure 6.15: Scaled L_{side} device process flow for gate-first (GF), lithographically-aligned (LA) MOSFET fabrication employing the "wrap-around" gate device layout and double patterning technique for ohmics, comprising three levels of EBL.

of gate dielectric in contact window, followed by metal deposition and lift-off. The ohmic anneal is carried out post large contact metallisation, in order to produce S/D contacts that are ohmic across both the stripe and large contacts.

6.4.3 Process Development

The ability to realise sub-100 nm L_{side} gaps requires a process of high resolution and contrast. This is best achieved with thin resist recipes, since the significance of electron scattering is considerably reduced, as discussed in Section 3.2.3.2. However, the minimum resist thickness that can be adopted in the process is limited by existing topography and ohmic metal thickness. The thickness of the resist should be chosen such that it provides sufficient coverage over existing topography. In addition, for metal lift-off to be effective, the resist needs to be twice as thick as the deposited metal. This however is more a rule of thumb, and in fact, it is possible to achieve good lift-off with thinner resists as long as the thickness of the metal is sufficiently smaller.

Based on the existing gate topography of 120 nm, a 4% 2010 (153 nm) / 2.5% 2041 (38 nm) PMMA bilayer resist seemed to be a suitable choice. However, the best performing Au/Zn/Au contacts to pre-MOSFETs had a thickness of 200 nm, this is incompatible for lift-off. A quick screening of thinner metal recipes on MOSFET material identified Au(5 nm)/Zn(35 nm)/Au(100 nm) to have contact performance comparable to the 200 nm thick metallisation on pre-MOSFETs. Therefore, the aforementioned bilayer resist was adopted for process investigation.

A sample, comprising of an InP substrate, was put through two levels of lithographic patterning for dose testing the new process. In the first level, gate patterns of 100 nm, 300 nm and 1 μm in length, with a fixed width of 100 μm, were patterned and metallised with Al(20 nm)/Au(100 nm). Subsequent to this 2 μm$\times$100 μm stripe patterns were aligned to the gates for a range of L_{side} gaps designed from 20 nm up to 1 μm. The patterns were then exposed in the VB6 tool at 100 kV using a beam current of 8 nA, with a corresponding spot size of approximately 12 nm. A dose range of $200 - 500\,\mu\text{C.cm}^{-1}$ was investigated at each gap spacing. Following exposure, the sample was developed in 2:1 mixture of IPA:MIBK for 1 min at 23 °C, and metallised with 140 nm thick Au/Zn/Au using a thermal evaporator. All metal deposition steps were preceded by oxygen plasma ashing to remove resist residues.

Figure 6.16 illustrates the scanning electron microscopy (SEM) profiles of device patterns with sub-100 nm gap designs. The smallest designs of 20 and 30 nm presented with metal lift-off issues. Furthermore, the stripe contacts are observed to overlap the gate at its edges, which would result in short circuits and device failure. This is a result of size bias [245] which is an effect caused by electron scattering of the resist, inherent to EBL. Therefore, the fabricated stripe contacts will have slightly larger dimensions compared to their design size, resulting in smaller than intended gap spacings. For gap designs $\geq$ 40 nm, lift-off was successful, however overlapping edges are still noticeable for 40 nm and 50 nm designs. In addition, there is also the problem of metal flagging at the inner edge of stripe contacts, which has the potential to cause device short circuiting. The aforementioned issues were not apparent at the 80 nm design and larger, which yielded well lifted-off stripe contacts, with clearly visible gaps.

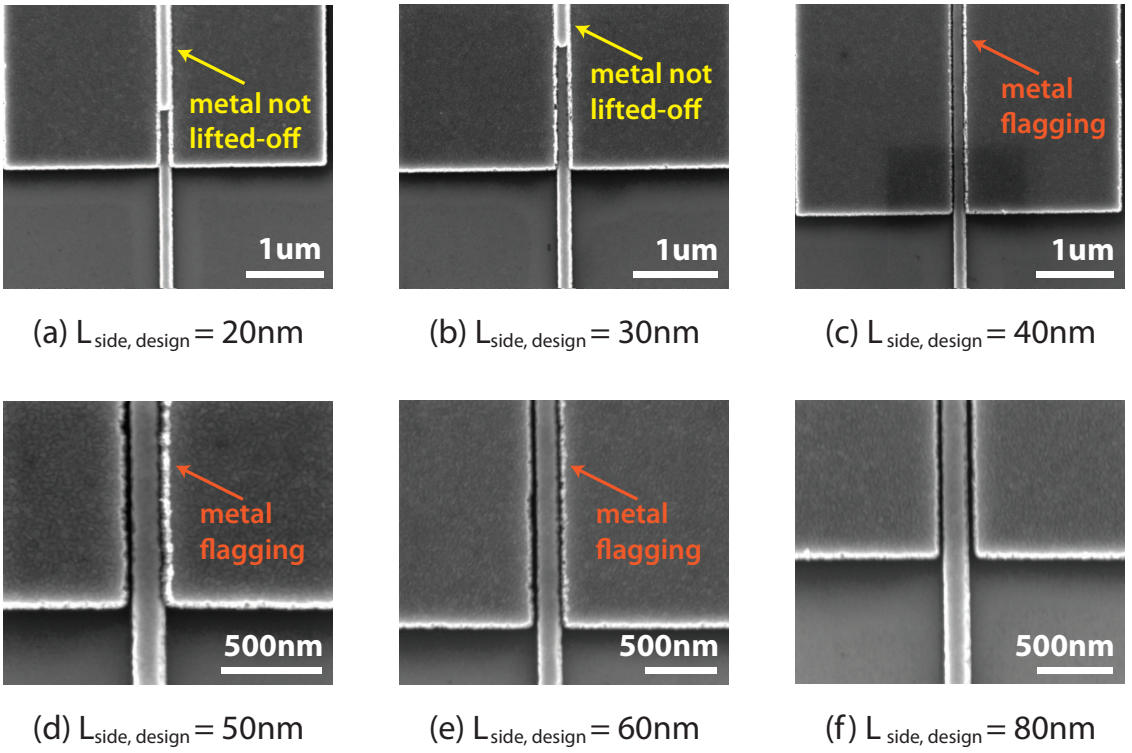


Figure 6.16: SEM profiles illustrating metal lift-off and flagging issues for sub-100 nm gap designs in device patterns.

It is worth noting physical gaps as small as 20 nm were realised with this process. However, the effect of line edge roughness was significantly higher at the inner edges of the stripe contacts at such gap dimensions. This results in a non-uniform gap along the width of the contact, likely translating to variability in the electrical performance of fabricated devices. It is believed the particular ohmic metallisation of Au/Zn/Au and its associated thickness are the causes of this increased edge roughness. For this reason gap designs ≤ 100 nm were deemed unsuitable for device implementation.

Although gaps of several sizes were investigated, the following discussion will be restricted to the gap designs implemented in the final device. Apart from the 1 μm gap spacing, already in use for baseline devices, three other gap designs of 120 nm, 260 nm and 520 nm were adopted. Table 6.4 is a summary of the effective exposure dose range at each of the gaps, based on the dose test results.

Table 6.4: Dose ranges for 120 nm, 260 nm and 520 nm gap designs.

Gap design	Effective dose range (μC/cm^2)
120 nm	230-290
260 nm	280-330
520 nm	230-330

It can be seen from Table 6.4 that the dose ranges provide for a reasonable process latitude at each gap. The resulting gaps for a specific dose within this effective range are shown in Figures 6.17, 6.18 and 6.19, for the 120 nm, 260 nm, and 520 nm designs respectively. From the SEM micrographs the gap sizes were measured using ImageJ [246], an open source image processing software. As before, measured gaps were smaller than the designed gap spacing. The size bias was consistent at $50-60$ nm for all three design gaps, across the gate lengths. Two observations are noted, first the gaps on either side of the gate were not identical. In addition, the gap spacing on the source side (left of the gate) was consistently higher than the drain side. These differences are due to the lithographic process, which can be further optimised to yield more consistent gap spacings.

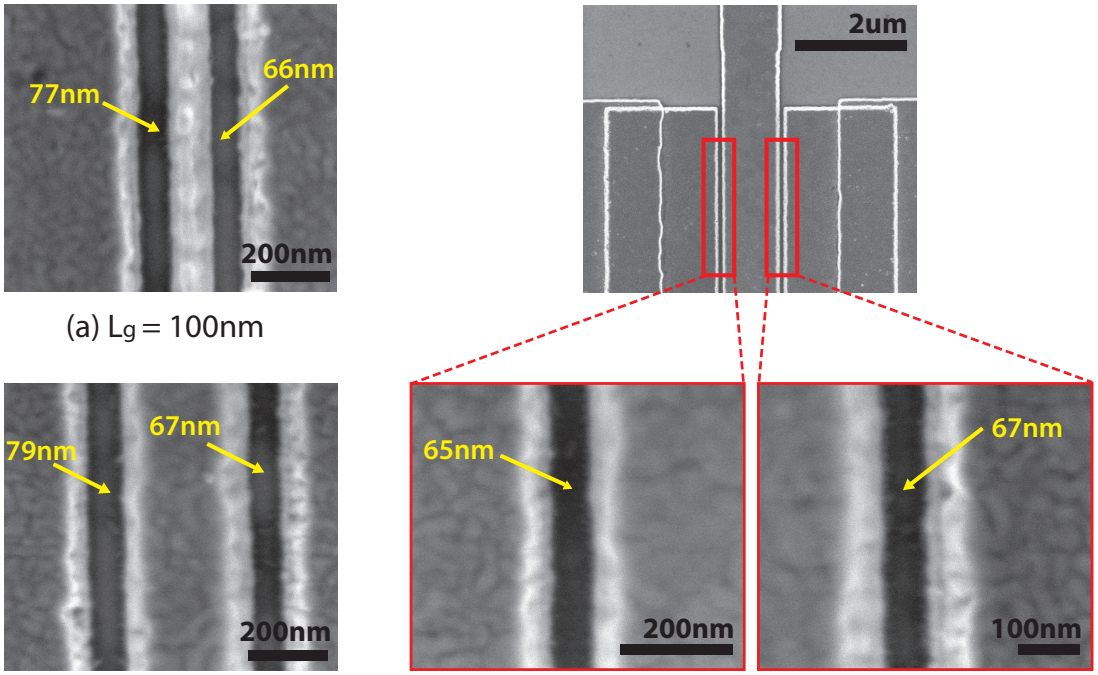


Figure 6.17: Plan view SEM micrographs depicting actual L_{side} for a design of 120 nm.

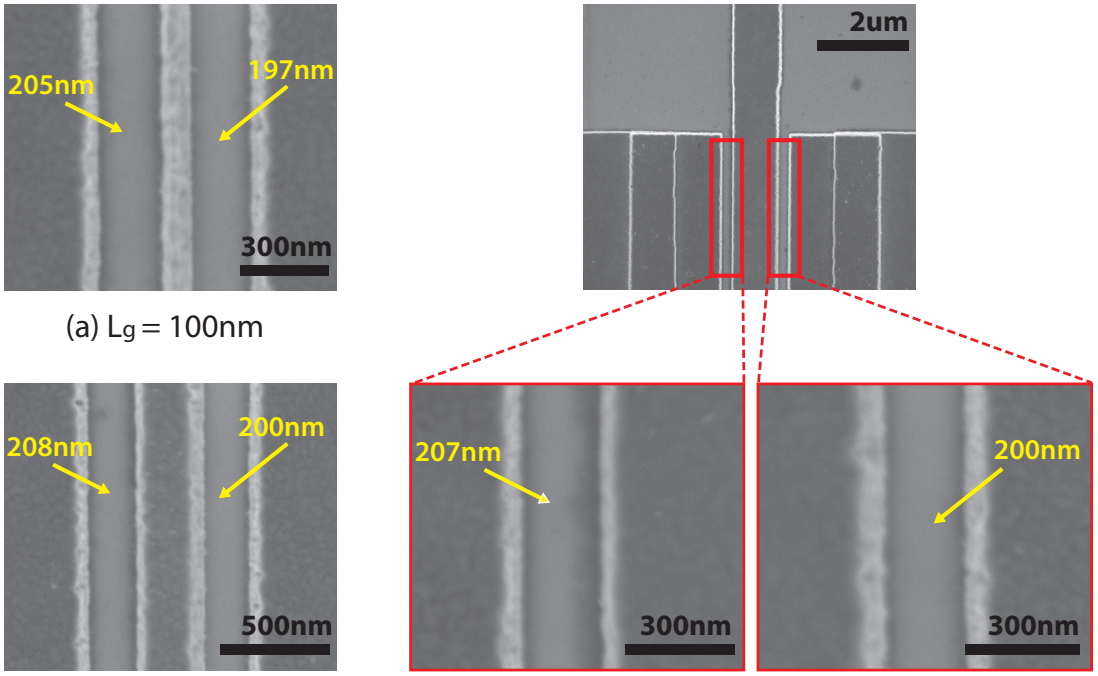
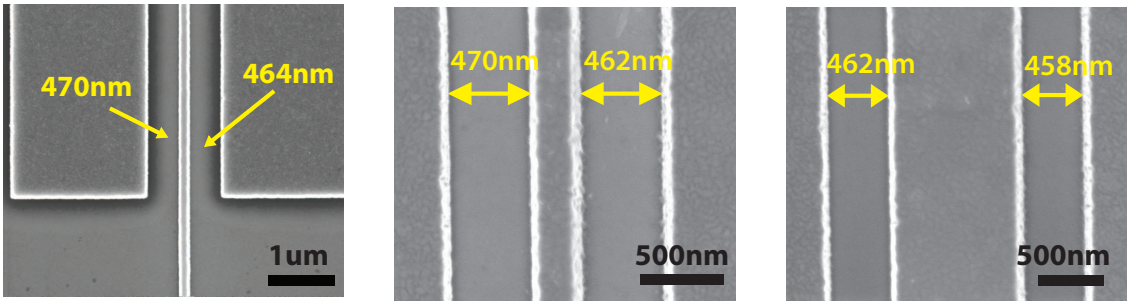


Figure 6.18: Plan view SEM micrographs depicting actual L_{side} for a design of 260 nm.



(a) Lg = 100nm (b) Lg = 300nm (b) Lg = 1um

Figure 6.19: Plan view SEM micrographs depicting actual L_{side} for a design of 520 nm.

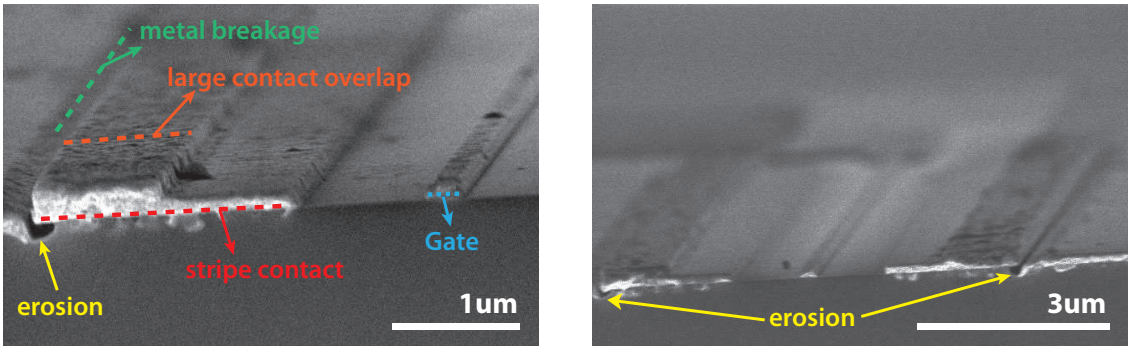
6.4.4 Process Qualification

With the processing parameters for scaled L_{side} device fabrication established, it was important to qualify the complete process flow (see Figure 6.15) on MOSFET material. For this purpose an InGaAs device sample, featuring a 6.5 nm thick Al_2O_3 gate dielectric, was used.

6.4.4.1 Process Flow Issues

A cross-section of the sample, upon which the complete device flow had been applied, was inspected under the SEM to mechanically validate the fabrication process. Based on the SEM profile of Figure 6.20(a), there appeared to be definite erosion of the substrate material at and around the outer edge of the stripe contact. Such erosion was found extending along the outer edges of stripe contacts in both the source and drain regions as illustrated in Figure 6.20(b). Furthermore, there was evidence of a break in the large contact at the outer edge of overlap with the stripe contact, coincident with the erosion site. Since current is forced through the large contacts during 3-terminal FET measurements, such metal discontinuity would result in an inoperable device and therefore needed resolving.

It was apparent the problem arose between stripe and large contact patterning. In order to identify the underlying cause, SEM profiling was done after each level of patterning, with the exception of gate definition. Cross-sectional inspection following stripe contact lithography, Al_2O_3 wet etch, metallisation and lift-off did not reveal any substrate erosion, as evidenced from Figure 6.21(a). A second sample, previously defined with stripe contact, was then subjected to the large contact lithography. Following only wet etching of the Al_2O_3 and resist removal in acetone, the sample was again inspected. This revealed substrate erosion (Figure 6.21(b)) identical to that observed in Figure 6.20. This



(a) Profile of source and gate regions (b) Profile of source/drain and gate regions

Figure 6.20: SEM profiles depicting semiconductor material erosion along the outer edge of stipe contact and a break in the large contact at the erosion site of devices fabricated using the scaled L_{side} process flow.

meant subsequent metallisation for large contact, rather than being continuous across the overlapping edge, would fall into the trough at the erosion site causing the previously highlighted breakage. On first thought, it seemed unlikely that the wet etching step was responsible for the erosion given that the etchant in question is a photoresist developer (AZ400K). If it was, such erosion should have also shown up during stripe contact patterning. There was one notable difference between the two levels of lithographic patterning, only the substrate material is exposed in the resist windows opened for stripe contact patterning. On the other hand, the resist template for large contact patterning exposes both the substrate and a portion of the stripe metal. This stems from the requirements of a 1μm overlap between the stripe and large contacts in the process flow. It is then possible the exposed metal could have affected the wet etching characteristics in a manner that causes erosion. Similar findings have been reported for wet treatments of semiconductor material with exposed metal in close proximity [247–251] and is ascribed to electrochemical etching effects.

Electrochemical etching is a chemically driven reaction involving the flow of charge between two electrodes, an anode and a cathode, of different surface potentials in an electrolytic solution [249]. During the wet etching step in the process flow given in Figure 6.15, the partially exposed stripe metal contact and the substrate to be etched behave as a cathode and anode respectively, while the dilute KOH forms the electrolytic solution. The established electrochemical circuit, and hence current will result in enhanced etch characteristics. A detailed description of the electrochemical etch and its associated effects/observations can be found in [250, 251].

There are a few methods in which the effect of the electrochemical etching can be

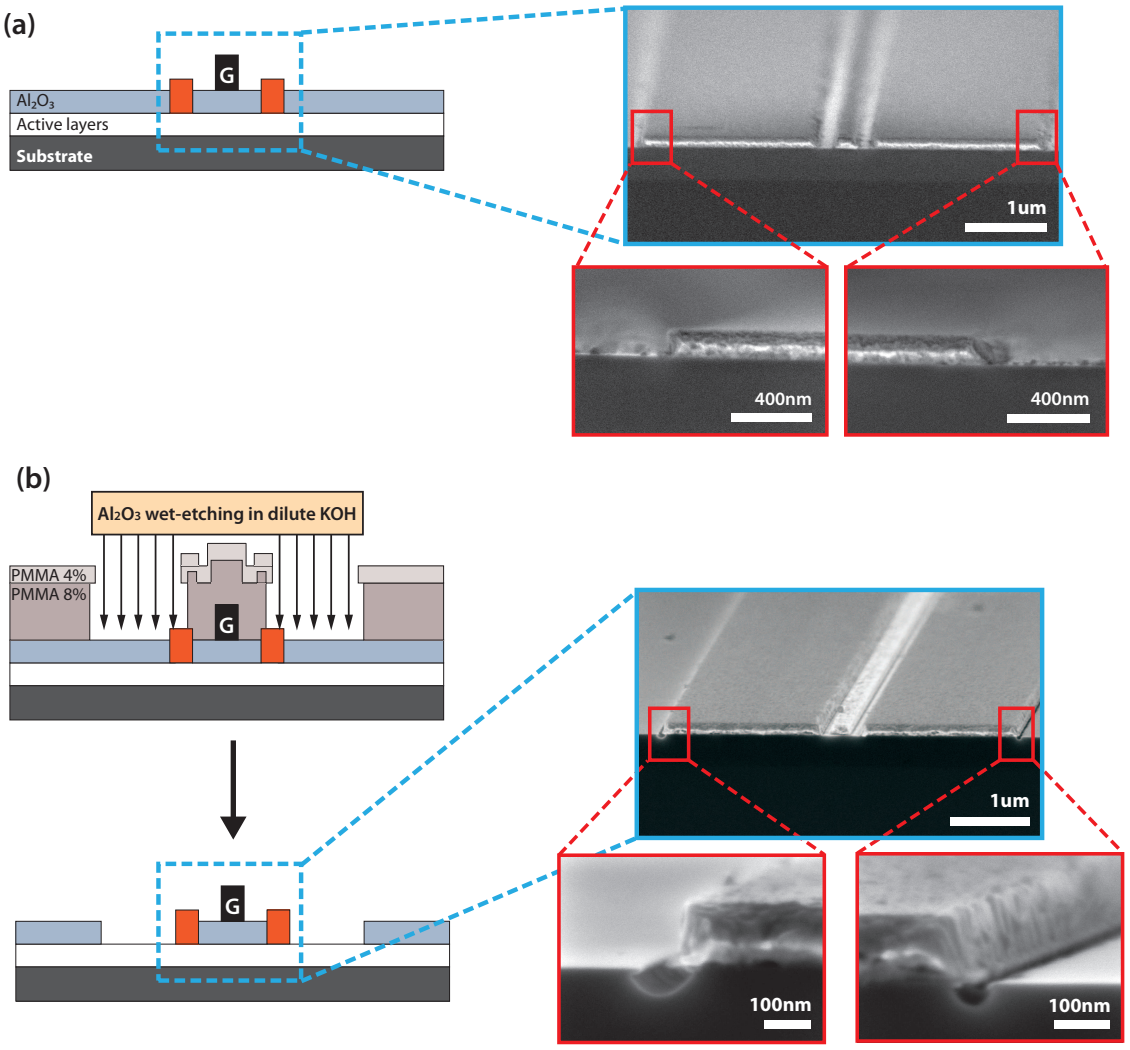


Figure 6.21: SEM inspection of device profile at different stages of scaled L_{side} process flow to identify the issue of material erosion: (a) after 2nd level EBL (stripe contact), Al_2O_3 wet etch, metallisation and lift-off, and (b) after 3rd level EBL (large contact), Al_2O_3 wet etch and resist removal in acetone.

suppressed [247–249]; increasing the proximity of the exposed ohmic metal to the substrate needing to be etched, reducing the size of the exposed metal or avoiding noble metals such as Au with higher electrode potential to minimise the excessive oxidation of semiconductors. However, none of these methods are feasible. An overlap between the stripe and large contacts as well as the AuZn-based ohmic scheme are key requirements of the process flow. Even if the size of the overlap could be reduced, this will still not mitigate the etch issue.

6.4.4.2 Revised Process Flow

Since the dimensions of the stripe contact was designed taking into consideration the transfer length of ohmic contacts on III-V epi used in this work, it was not necessary for the large contacts to be actually ohmic in nature. The large contacts however needed to be in place serving as a bondpad to provide a means to device measurement. In view of this the Al_2O_3 wet etch following large contact lithography was removed from the flow. Moreover, the Au/Zn/Au metallisation for large contact patterning was replaced with Ti/Pt/Au. A final change to the process flow was to move the ohmic anneal, initially carried out after large contact metallisation, to after stripe contact metallisation. The process flow incorporating these changes is illustrated in Figure 6.22. This revised flow was adopted for scaled L_{side} device fabrication and the results of this will be discussed in Chapter 7.

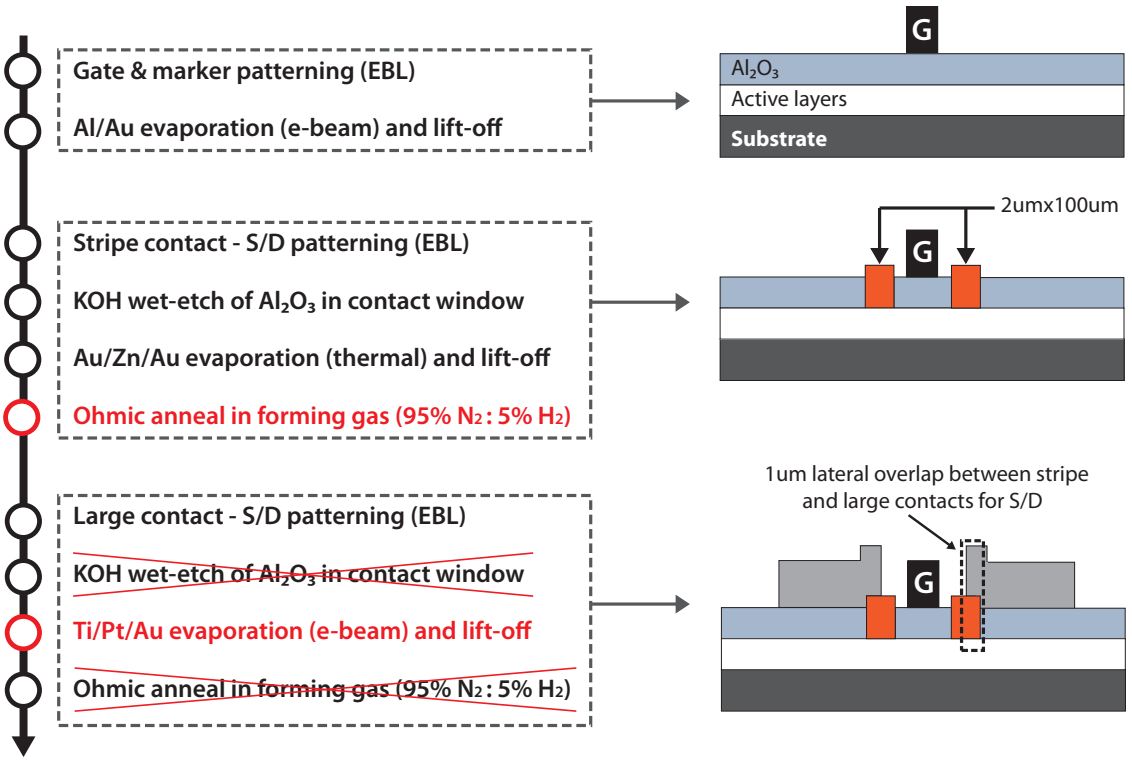


Figure 6.22: Modified process flow for scaled L_{side} device fabrication. Changes from the previous flow are marked in red.

6.4.5 Summary

Large R_{side} between the S/D contacts and the gate can significantly limit the performance of MOSFETs. One component of R_{side} arises from the gate-to-source/drain gap, L_{side}.

Scaling this gap into the deep submicron region would yield increased $I_{\mathrm{d,sat}}$, reduced R_{sd} and may promote non-equilibrium high-field transport at short L_{g}.

Both SA and LA techniques were considered for L_{side} gap scaling. A LA technique based on the double ohmic patterning approach was chosen for investigation, since it was found to be process-compatible with the III-V p-MOSFETs of this work. The resulting device layout, and process flow for implementing this approach, was presented. The process was developed through dose tests and SEM inspection. A highlight of the new fabrication process is the realisation of gaps as small as 20 nm.

A mechanical validation of the complete flow was carried out on a relevant substrate material. From SEM interrogation, a critical issue - electrochemical etch effect - with the process flow was identified. This was remedied by revising the process flow, requiring only minimal changes and no additional process development.

Based on the potential for significant performance enhancement in devices, the process was employed on a InGaAs-channel device wafer. The results of this will be presented in the next chapter.

6.5 High-k/GaSb Interface Engineering

6.5.1 Introduction

Gate stack modules comprising high-k gate dielectrics integrated on InGaAs and GaSb were required for device realisation in this work. As high-k/InGaAs gate stack modules, available within the Nonclassical CMOS Research Centre, were mature, no further development was needed. Focus, therefore, was placed on high-k integration on GaSb.

The higher hole mobilities arising from strained, antimony-based channels offer a means to realising a high performance, low power p-MOS solution. However, achieving a low-defect, high-quality dielectric/semiconductor interface remains the most notable impediment to implementing a III-V logic solution. This is of even greater concern to antimonides given their higher inherent susceptibility to ambient air exposure [208]. As a result antimony surfaces are terminated with native oxides that are neither stable, self-limiting nor abrupt [209, 252]. The increase in root-mean-square (RMS) roughness of GaSb substrates from 0.73 nm to 4.01 nm following air exposure for a week, as reported by Nainani *et al.* [209], testifies to such native oxide growth. The resulting defect-dominated interface impairs E_{F} movement, which in turn limits the MOSFET's ability to effectively

modulate the channel charge [253]. Therefore, the advantage of incorporating an antimony channel would only go so far in improving device performance, if the quality of the high-k/antimony gate stack is sub-optimal. In this work, antimony-based p-MOSFETs are realised on device epi terminated with GaSb. This then motivates an investigation of a suitable passivation scheme to engineer a high-quality, low-defect high-k/GaSb interface to realise the envisaged performance benefits of antimonide-based p-MOSFETs.

A variety of surface passivation approaches, prior to gate dielectric deposition using a number of techniques including ALD, MBE and electron beam evaporation (EBE), have been explored as means to removing the native oxide on GaSb(100). These approaches comprise wet chemical etchants, hydrogen plasma cleaning, interface passivation layers (IPL) and high-temperature thermal desorption (see Table 6.5). Of the wet chemical approaches, treatments based on HCl have been found to be effective in minimising surface oxides, resulting in high-k/GaSb interfaces with improved electrical properties [85, 93, 209, 254]. On the other hand, ammonium sulphide ($(NH_4)_2S$), a wet treatment that has shown to engineer a high-quality high-k/InGaAs [26, 37, 255, 256] resulting in midgap D_{it} as low as 5×10^{11} cm^{-2}eV^{-1} [256], has received little attention on GaSb. One study [257] reported the removal of Sb oxides after 2% sulphide treatment. The XPS analysis also revealed the Ga oxide content to be lower for $(NH_4)_2S$ compared to NH_4OH or HCl treatments. This in conjunction with the absence of Sb oxides resulted in a larger capacitance swing in the C–V response of the sulphide treated sample. The removal of Sb oxides and the retention of Ga oxides were also noted for 22% sulphide treatments [258]. The effects of the treatment however were not electrically evaluated in the study. Other electrical examinations have been limited to combined treatments of $(NH_4)_2S$ and HCl [259–263]. The fact remains that presently a systematic investigation of the impact of $(NH_4)_2S$ on the high-k/GaSb interface for sulphide concentrations in the range 1-22%, similar to that reported for InGaAs [26], is lacking. Therefore, the effectiveness of $(NH_4)_2S$ as a standalone surface treatment for Al_2O_3/p-GaSb MOS system is investigated in this work. Frequency dependent C–V measurements are used to assess the effects of the treatment for varying sulphide concentrations.

There are two significant differences between III-As and III-Sb surfaces exposed to $(NH_4)_2S$ treatments. Interactions of sulphur with III-As results in the formation of a monolayer (ML) thick sulphide layer [264]. In addition, such treatments performed at RT hardly etch the III-As surfaces [26]. It can then be said that sulphur is chemically active but physically inert on III-As semiconductors. In contrast III-Sb surfaces exposed to similar treatments present with thicker sulphide layers and undergo considerable etching.

Table 6.5: Summary of interface passivation techniques and resulting interface trap densities (D_{it}) reported in literature for high-k/GaSb (100) MOS capacitors. Acronyms used: atomic layer deposition (ALD), electron beam evaporation (EBE), molecular beam epitaxy (MBE), plasma-enhanced ALD (PEALD); de-ionised water (DIW), forming gas (FG), interface passivation layer (IPL), non-intentionally doped (NID), post deposition anneal (PDA), post metallisation anneal (PMA), ultra-high vacuum (UHV); bandgap (BG), conduction band (CB), midgap (MG), valence band (VB).

Doping	High-k/deposition method	Passivation technique	D_{it} analysis method	D_{it} ($\mathrm{cm^{-2}\,eV^{-1}}$)	Refs.
n-type	Al_2O_3 ALD (300°C)	HCl	Conductance	4.5×10^{13} (VB edge), $\sim 3.8 \times 10^{13}$ (MG & upper BG)	[265]
		HCl, FG PMA		2.5×10^{13} (VB edge), 2.9×10^{13} (CB edge), 2.1×10^{13} (MG)	
n/p-type	Al_2O_3 ALD (300°C)	HCl	Modelling/ Conductance	1.8×10^{13} (VB edge), $0.9{-}1 \times 10^{13}$ (MG & upper BG)	[259, 260]
		HCl, $(NH_4)_2S$		1.6×10^{13} (VB edge), 1.3×10^{13} (MG), 5×10^{12} (CB edge)	
	Al_2O_3 PEALD (200°C)	HCl		1.5×10^{12} (VB edge), 5.5×10^{12} ($\sim$MG), 9×10^{12} (upper BG)	
		HCl, $(NH_4)_2S$		2×10^{12} (VB edge), 9×10^{12} ($\sim$MG), $1{-}1.2 \times 10^{13}$ (CB edge)	
p-type	HfO_2 ALD (250°C)	in-situ H_2 plasma, in-situ FG PDA	Terman	$\geq 1.3 \times 10^{14}$ (close to VB)	[40]
		HCl, in-situ H_2 plasma, in-situ FG PDA		1.8×10^{14} (VB edge), $0.5{-}3 \times 10^{14}$ (lower BG)	
n-type	Al_2O_3/HfO_2 EBE (25°C)	α−Si IPL, FG PDA	Conductance/ High-low/Terman	$0.35{-}1.5 \times 10^{13}$ (upper BG)	[91]
p-type	HfAlO ALD (200°C)	Hf-1st	Conductance	$4.1{\times}10^{12}$ (VB edge), 3.8-$4.5{\times}10^{12}$ (lower BG)	[266]
		Al-1st		$4{\times}10^{12}$ (VB edge), 4-$6{\times}10^{12}$ (lower BG)	
		Hf-1st, N_2 PDA		$1{\times}10^{13}$ (VB edge), 1-$1.8{\times}10^{13}$ (lower BG)	
		Al-1st, N_2 PDA		$8{\times}10^{12}$ (VB edge), 0.8-$1.2{\times}10^{13}$ (lower BG)	

continued....

Doping	High-k/deposition method	Passivation technique	D_{it} analysis method	D_{it} ($cm^{-2}\,eV^{-1}$)	Refs.
n-type	Al_2O_3 ALD (300°C)	HCl	High-low	$0.7-2\times10^{13}$ (lower BG), 7×10^{12} (MG), $0.65-1.5\times10^{13}$ (upper BG), 1.5×10^{13} (CB edge)	[267]
		2 nm n-InAs surface layer, HCl	High-low	$0.2-1.7\times10^{13}$ (lower BG), 2.1×10^{12} (MG), $1.1-2.1\times10^{12}$ (upper BG), 1.35×10^{12} (CB edge)	
		2 nm n-InAs surface layer, HCl, $(NH_4)_2S$, FG PDA	Conductance/ High-low/Terman	$0.6-0.7\times10^{12}$ (lower BG), $0.4-0.8\times10^{12}$ (MG), $0.06-2\times10^{13}$ (upper BG)	
	HfO_2/Al_2O_3 ALD (300°C)	2 nm n-InAs surface layer, HCl, $(NH_4)_2S$, FG PDA	Conductance/ High-low/Terman	$0.15-1.4\times10^{12}$ (lower BG), $0.95-1.4\times10^{12}$ (MG), $0.095-2\times10^{13}$ (upper BG)	
n/p-type	Al_2O_3 ALD (300°C)	HCl, FG PDA	Conductance	$0.1-1\times10^{12}$ (lower BG), 3×10^{11} (MG), $0.3-15\times10^{12}$ (upper BG)	[85, 93]
n-type	HfO_2/Al_2O_3 ALD (200°C)	HCl, unneutralised $(NH_4)_2S$	Conductance	2×10^{12} (MG), $2-3\times10^{12}$ (upper BG closer to MG)	[261]
		HCl, neutralised $(NH_4)_2S$		$1.95-2.8\times10^{12}$ (upper BG closer to MG)	
p-type	Al_2O_3 PEALD (150°C)	FG PDA	Terman	$\geq1.85\times10^{14}$ (close to VB)	[268]
		HCl, FG PDA		$\geq0.85\times10^{14}$ (close to VB)	
		in-situ H_2 plasma, TMA pre-pulsing, FG PDA:			
		1. H_2 - 50 W, 10 min, 250°C		1.5×10^{13} (MG), $0.9-1.5\times10^{13}$ (lower BG)	
		2. H_2 - 50 W, 30 min, 150°C		3×10^{12} (MG), $2.5-9\times10^{12}$ (lower BG)	
		3: H_2 - 100 W, 10 min, 150°C		$\sim0.5\times10^{12}$ (MG), $0.5-5.5\times10^{12}$ (lower BG)	
p-type	Al_2O_3 ALD (300°C)	HF, UHV anneal	Conductance	5×10^{12} (MG)	[209]
		NH_4OH, UHV anneal		...	
		HCl, UHV anneal		3×10^{11} (MG)	

continued....

Doping	High-k/deposition method	Passivation technique	D_{it} analysis method	D_{it} ($cm^{-2}\,eV^{-1}$)	Refs.
p-type	HfO_2 PEALD (150°C)	FG PDA ex-situ H_2 plasma, TMA pre-pulsing, FG PDA	Conductance	$\geq 5 \times 10^{14}$ (lower BG close to MG) $\sim 1 \times 10^{13}$ (lower BG close to MG)	[269]
	HfO_2/Al_2O_3 PEALD (150°C)	ex-situ H_2 plasma, TMA pre-pulsing, FG PDA		$\sim 1 \times 10^{13}$ (lower BG close to MG)	
p-type	Al_2O_3 PEALD (150°C)	in-situ H_2 plasma, TMA pre-pulsing, FG PDA	High-low	$\leq 1 \times 10^{13}$ (MG)	[253]
p-type	Al_2O_3 MBE (50°C)	As decap in UHV, FG PDA, FG PMA	Conductance/ Terman	2×10^{12} (MG), $1.5-4.3 \times 10^{12}$ (lower BG)	[270]
n/p-type	Al_2O_3 ALD (320°C)/ RT-Y_2O_3 MBE Al_2O_3 ALD (320°C)/ HT-Y_2O_3 MBE	As decap in UHV, FG PDA	Conductance/ Gray-Brown	$\sim 1 \times 10^{12}$ (lower BG), $2-8 \times 10^{13}$ (upper BG) $1.8-3.5 \times 10^{12}$ (upper BG)	[87, 271]
n-type	Al_2O_3 ALD (300°C)	PDA HCl, PDA HCl, UHV anneal, PDA	Conductance	$\geq 1 \times 10^{14}$ ($\sim$MG) $\sim 1 \times 10^{13}$ ($\sim$MG) $\sim 1 \times 10^{12}$ ($\sim$MG)	[272]
p-type	Al_2O_3 ALD (150°C) Al_2O_3 ALD (200°C) Al_2O_3 ALD (250°C) Al_2O_3 ALD (300°C)	cyclic HCl	Terman	$\geq 4 \times 10^{13}$ (close to VB), 5.5×10^{13} (VB edge) $\geq 5 \times 10^{13}$ (close to VB), 7×10^{13} (VB edge) $\geq 5 \times 10^{13}$ (close to VB), 8×10^{13} (VB edge) $\geq 7 \times 10^{13}$ (close to VB), 8×10^{13} (VB edge)	[90, 273, 274]
NID	HfO_2/Al_2O_3 ALD (200°C) HfO_2/Al_2O_3 ALD (110°C)	HCl	Modelling	$\sim 2 \times 10^{13}$ (VB edge) $\geq 6 \times 10^{12}$ (lower BG), $\sim 7 \times 10^{12}$ (VB edge)	[275]

continued....

Doping	High-k/deposition method	Passivation technique	D_{it} analysis method	D_{it} ($\text{cm}^{-2}\,\text{eV}^{-1}$)	Refs.
n/p-type	Al_2O_3 ALD (270°C)	HCl, UHV anneal, TMA pre-pulsing, FG PMA	···	$6-7\times10^{11}$ (lower BG), $2.5-4\times10^{12}$ (upper BG)	[276]
		HCl, UHV anneal, TMA+ H_2O pre-pulsing, FG PMA		$\sim5\times10^{11}$ (lower BG), $2-3\times10^{12}$ (upper BG)	
n-type	HfO_2 ALD (200°C)	HCl, unneutralised $(NH_4)_2S$	Modelling	$1.5-2.4\times10^{13}$ (upper BG closer to MG)	[262]
		HCl, neutralised $(NH_4)_2S$		$1.2-2.2\times10^{13}$ (upper BG closer to MG)	
n/p-type	Al_2O_3 ALD (300°C) †	HCl, NH_4OH, N_2 PDA	Conductance/ High-low	2×10^{12} (VB edge), 2×10^{13} (CB edge), 2.5-4$\times10^{13}$ (upper BG)	[84]
	Al_2O_3 ALD (300°C) ‡			2×10^{12} (VB edge), 2-3$\times10^{13}$ (upper BG)	
	Al_2O_3 ALD (200°C) ‡			2×10^{12} (VB edge), 1.5-2$\times10^{13}$ (upper BG)	
p-type	Al_2O_3 ALD (300°C)	HCl, $(NH_4)_2S$, TMA pre-pulsing	Modelling	$0.2-2\times10^{13}$ (upper BG), 2×10^{13} (CB edge)	[263]
NID	Al_2O_3 ALD (190°C)	no treatment	Fitting	$3.3-4.1\times10^{13}$	[254]
		HCl (no DIW rinse)		$<1\times10^{12}$	
		HCl (DIW rinse)		$6.3-8.1\times10^{13}$	
n/p-type	Al_2O_3 EBE (25°C)	FG PDA	Terman	2×10^{13} (VB edge), $0.1-2\times10^{13}$ (lower BG),	[277]
		α−Si IPL, FG PDA		$\sim0.4\times10^{13}$ (MG)	
	Al_2O_3/HfO_2 EBE (25°C)	α−Si IPL, FG PDA		$3.5-8\times10^{12}$ (upper BG), 8×10^{12} (CB edge)	

† Dopant activation at 650 °C rapid thermal anneal (RTA)
‡ Dopant activation at 600 °C RTA

InSb exposed to $(NH_4)_2S$ for 60 min at RT results in a sulphide layer that is 6-7 ML thick [264]. Sulphide layers formed on GaSb are much thicker by nature [278]. Moreover, etch rates of sulphur-based treatments on GaSb are considerably higher [279]. This then motivates a study of the surface topography for determining the roughness and etch characteristics resulting from sulphide treatments.

In the following discussion, details of the grown epi-layer as well as the $(NH_4)_2S$ passivation conditions used in the investigations are highlighted. The section concludes with the analysis and discussion of the experimental results.

6.5.2 Experimental Details

The p-doped GaSb wafer used in this study, as depicted in Figure 6.23, was grown at Texas State University. It employed a heavily doped (Zn: 0.8–1.4×10^{19} cm^{-3}) GaAs(100) substrate on which a 1 μm AlSb buffer (Be: 5×10^{18} cm^{-3}) was grown. The stack was completed with 150 nm GaSb (Be: 5×10^{18} cm^{-3}) and 500 nm GaSb (Be: 4×10^{17} cm^{-3}) layers.

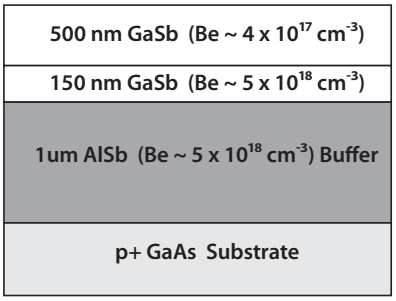


Figure 6.23: Cross-section of p-doped GaSb/AlSb/GaAs C–V wafer stack, 7-144.

Surface treatment and subsequent gate dielectric deposition were performed at the Tyndall National Institute. The GaSb surfaces were degreased sequentially in acetone, methanol and IPA at RT for 1 min each, and dried with flowing N_2 prior to the treatment. The ex-situ treatment conditions comprised $(NH_4)_2S$ diluted to differing concentrations in de-ionised water (DIW). Two aspects of the treatment were addressed during the study; the effectiveness of sulphide as an interface passivant and its impact on the surface topography of GaSb.

Apart from the concentration of the sulphide solution, a number of other process parameters can also have a strong bearing on both aspects of the treatment highlighted earlier. These include any pre-treatments undertaken prior to the sulphide treatment, the temperature of the sulphide solution, and sample immersion time (passivation time) in

the solution [26]. In this work $(NH_4)_2S$ was adopted as a standalone treatment performed at RT (∼295 K) with no prior surface pre-treatments, apart from a degrease step. The immersion time was kept fixed for the different sulphide concentrations in the electrical passivation part of the study. In contrast, several immersion times were investigated in conjunction with varying sulphide concentrations during the surface topography study.

The efficacy of $(NH_4)_2S$ as a passivant was assessed on GaSb epi-layers subjected to different surface treatments prior to gate dielectric deposition. This was investigated for four different chemical treatments as summarised in Table 6.6. The first sample analysed was not subjected to any treatment (native surface), and served as a control in the experimental study. Samples (b)-(e) were treated with 1%, 5%, 10% and 22% aqueous $(NH_4)_2S$ solutions respectively, for which immersion time in the solution was kept constant at 10 min. The samples were then loaded into the ALD system with a lag time of ∼4 min between removal from the aqueous solution and the loading; minimum air exposure of samples ensures ambient contamination as well as native oxide re-growth are minimised [26]. A 8 nm-thick (nominal) Al_2O_3 film was deposited by ALD using alternating pulses of TMA and H_2O at 300 °C, in a TMA-first sequence.

Table 6.6: Summary of surface treatments performed on p-GaSb samples (7-144), prior to gate dielectric deposition for MOS capacitor realisation and characterisation.

Sample	Description
(a)	No surface treatment
(b)	1% $(NH_4)_2S$ @ 10min
(c)	5% $(NH_4)_2S$ @ 10min
(d)	10% $(NH_4)_2S$ @ 10min
(e)	22% $(NH_4)_2S$ @ 10min

The structural integrity, physical gate dielectric thickness (t_{ox}) and the presence of any interfacial layer (IL) between Al_2O_3 and GaSb were determined via cross-sectional TEM at Tyndall National Institute for select samples. Sample cross-sections for TEM were prepared using focused ion-beam (FIB) thinning procedures in a FEI Quanta 3Di workstation and analysed at 200 kV in a JEOL2100 system. MOS capacitor fabrication, and subsequent electrical characterisation, were undertaken at the University of Glasgow. Ni(60 nm) and Au(80 nm) were sequentially deposited via e-beam evaporation through a shadowmask for the gate electrode. For a reliable comparative study across all samples, circular capacitors of 100 μm in diameter were used for electrical characterisation

performed under a light-tight and electrically-shielded condition.

GaSb surface topography following sulphidation was systematically investigated for a variety of treatments. While the concentrations of the sulphide solution were kept identical to the aforementioned passivation study, sample immersion times of 1, 5 and 10 min were investigated at each concentration, resulting in a total of twelve different treatments. Atomic force microscopy (AFM) was used to analyse the resulting surface topography in terms of RMS roughness and thickness of GaSb etched, since $(NH_4)_2S$ is known to have a much stronger etching effect on GaSb as opposed to InGaAs. In order to accommodate the latter analysis, samples were patterned using EBL to open windows in PMMA resist prior to the chemical treatments. Following resist removal in acetone, the thickness of the etched layer was elucidated via step height measurements at the boundary between the treated and untreated areas. Knowledge of the etch is particularly important since this needs to be taken into account during the growth of InGaSb-channel device layers which are terminated with a GaSb cap layer.

6.5.3 Results and Discussion

6.5.3.1 Interface Passivation Study

Room temperature multi-frequency (1 kHz to 1 MHz) C–V characteristics of samples without any treatment (control) along with 1%, 5%, 10% and 22% $(NH_4)_2S$ treatments, for an immersion time of 10 min, are illustrated in Figure 6.24(a)-(e) respectively. All the samples exhibit modulation of the capacitance with applied V_g, with the level of capacitance modulation being dependent on the concentration of the sulphide treatment prior to the ALD process for Al_2O_3 growth. The C–V response is clearly improved for the treatment comprising 1% $(NH_4)_2S$, which indicates a reduced D_{it}. However a further increase in the sulphide concentration results in the degradation of the C–V response. In the case of the 22% treated sample a significant reduction in the capacitance modulation with gate bias is noted. This is indicative of FLP at the Al_2O_3/GaSb interface, which suggests the movement of the Fermi level in the semiconductor is severely hampered by the presence of a large number of trap states.

Capacitance Swing

One measure of the charge modulation capability, and hence Fermi level movement, of the MOS capacitor is a comparison between the measured and ideal capacitance swing with gate bias. This is quantified through the capacitance modulation efficiency (CME)

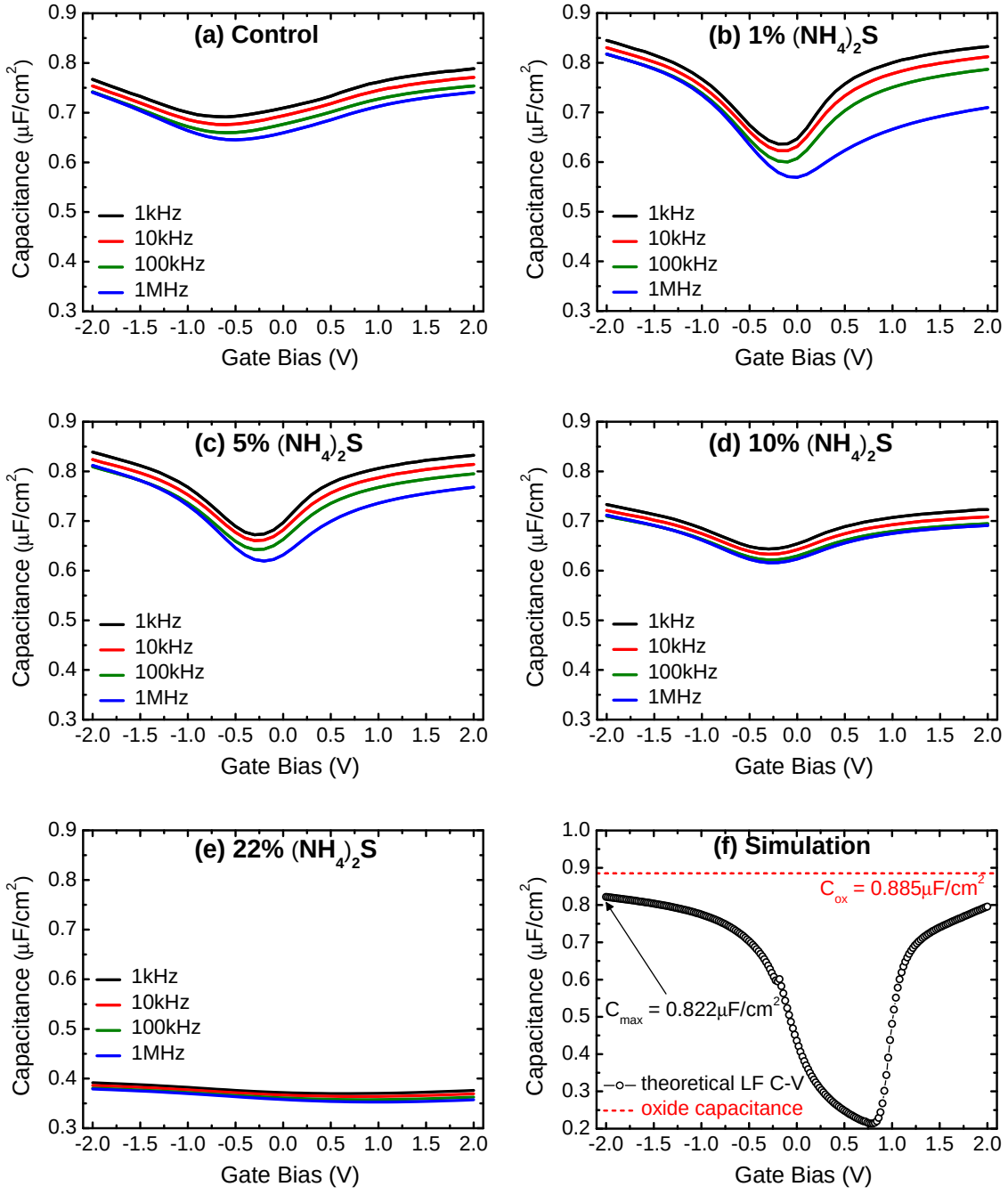


Figure 6.24: Multi-frequency (1 kHz to 1 MHz) C–V characteristics ($\sim$ 295 K) of (a) control (untreated) and (b) 1%, (c) 5%, (d) 10% and (e) 22% $(NH_4)_2S$ treated Au/Ni/Al_2O_3/p-GaSb MOS capacitors. The theoretical low frequency C–V characteristic calculated from one dimensional self-consistent Poisson-Schrödinger simulation is illustrated in (f) along with a marking representing the oxide capacitance.

given as:

$$\text{CME} = \frac{\left(\frac{C_{\max} - C_{\min}}{C_{\max}}\right)_{meas}}{\left(\frac{C_{\max} - C_{\min}}{C_{\max}}\right)_{ideal}} \times 100\,\%\,, \tag{6.5}$$

where $C_{\max}$ represents the maximum capacitance in the accumulation region and minimum capacitance ($C_{\min}$) is obtained between the regions of depletion and inversion. These capacitances are determined from C–V data measured at high frequencies. As will be explained later, the minority carrier response deteriorates at high frequencies, which in turn causes the capacitance in inversion to reach its minimum value. This therefore enables the maximum capacitance swing to be attained within the measured gate bias range spanning from accumulation to inversion. It should be noted that the terms "accumulation", "depletion" and "inversion" are used here purely for simplicity of expression to represent the nominal regions of the C–V response. The determination of genuine surface accumulation and inversion requires further analysis and will be discussed later in this section. The normalised capacitance swings of all samples, at a measurement frequency of 1 MHz, is shown in Figure 6.25. Also, demarcated is the normalised ideal capacitance swing. A one-dimensional self-consistent Poisson-Schrödinger (1D-PS) C–V simulation [280] is used in determining $C_{\max}$ and $C_{\min}$ for the ideal swing. In the simulation a dielectric constant of 8 [26] and the physical t_{ox} (based on TEM) are used for Al_2O_3.

The resulting CME of all samples is illustrated in Table 6.7. It should be pointed out that a CME of 100% relates to ideal charge modulation, which allows the Fermi level to

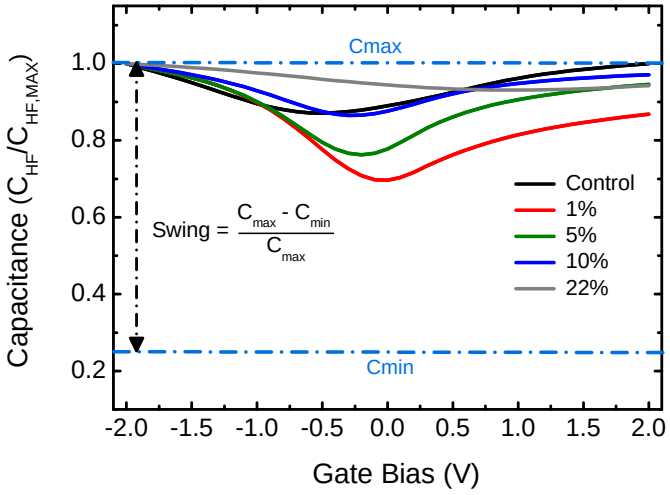


Figure 6.25: Normalised capacitance swings of control (untreated) and 1%, 5%, 10% and 22% $(NH_4)_2S$ treated Al_2O_3/p-GaSb MOS capacitors, determined from the 1 MHz C–V characteristic. Normalised ideal capacitance swing, based on 1D-PS C–V simulation, is marked as $C_{\min}$ and $C_{\max}$.

be completely swept from the conduction band through to the valence band. Based on this, the capacitance swing with gate bias appears fairly limited for all samples, with the highest CME of 40.6 % obtained for the 1% treated samples. This is indicative of a large number of surface traps obstructing the free movement of the Fermi level in the bandgap (E_g). The 1% treated sample shows a factor of two improvement in CME from the control sample, highlighting the effectiveness of the treatment in reducing D_{it}. However, the improved interface quality resulting from the sulphide treatment diminishes at higher concentrations as the charge modulation efficiency reduces. The rather small CME of 9.3% is further proof of FLP in the 22% treated sample.

Table 6.7: Summary of charge modulation efficiency (%), determined from the 1 MHz C–V response, of Al_2O_3/GaSb samples for different surface treatments.

Treatment	CME (%)
untreated	17.3
1% $(NH_4)_2S$	40.6
5% $(NH_4)_2S$	31.7
10% $(NH_4)_2S$	18.0
22% $(NH_4)_2S$	9.3

Inversion Response

As discussed in Section 2.3.3, the C–V characteristic in inversion is dependent upon the measurement frequency of the applied AC signal and τ_{min} of the semiconductor. So long as the period of the AC signal is much longer than τ_{min} ($1/\omega \gg \tau_{min}$), minority carriers will completely follow the applied signal. If this condition is satisfied, the capacitance in strong inversion saturates at a maximum value that approaches C_{ox}. Such an admittance response is referred to as LF behaviour. As the period of the AC signal reduces with increasing measurement frequency, the ability of the minority carriers to respond to the signal deteriorates. As a result, the carriers only partially follow the signal at these intermediate frequencies, resulting in a frequency-dependent C_{inv}. When the minority carriers are no longer able to follow the signal ($1/\omega < t_{min}$), C_{inv} saturates at a minimum value given by the series combination of C_{ox} and C_d (see Equation (2.34)). This characteristic represents the HF behaviour.

On inspecting the C–V responses in Figure 6.24, a notable feature is immediately observed in the inversion region. The control sample along with the 1%, 5% and 10% treated samples do not exhibit a HF behaviour, even at a measurement frequency of

1 MHz. The responses were more characteristic of LF and intermediate frequency behaviour, typical of minority carrier response, in the range of measured frequencies. Similar C–V responses are typically observed in high-k/III-V MOS capacitors featuring either InAs or InSb [281, 282]. The minority carrier response times of these III-Vs tend to be extremely short. For instance InSb has a $\tau_{\min}$ of 39 ns at RT [282]. This stems from the small bandgap (0.17 eV) [68], short τ_{L} (50 ns) and high n_{i} ($2\times10^{16}\,\mathrm{cm}^{-3}$) [282] of InSb. Analogous admittance response have also been reported for high-k/p-GaSb capacitors in literature for a variety of passivation schemes [84, 85, 93, 209, 259, 260, 266, 267]. This appears to suggest that the minority carrier response time of GaSb is possibly shorter than 1 μs.

Although GaSb has a short τ_{L} of 100 ns [283], it has an E_{g} of $\sim$0.726 eV and more importantly, an intrinsic carrier density that is four orders of magnitude smaller than in InSb ($n_{\mathrm{i}} = 1.405\times10^{12}\,\mathrm{cm}^{-3}$) [284]. This suggests that GaSb has a response time that is in fact much longer than 1 μs, and therefore should present with a HF response in the measured frequency range. In evidence to this, typical inversion characteristics presenting with LF, intermediate frequency and HF responses have been demonstrated on p-GaSb MOS capacitors [209, 253, 270]. It is therefore noted that the 1 MHz curve presents with a false inversion response arising from the high density of interface traps in the upper half of E_{g}.

Dispersion in Accumulation, Vfb Shift and Stretch-out

A number of metrics are of importance in assessing the quality of the high-k/III-V interface, as discussed in Section 4.4.1.2. These include stretch-out, flatband voltage (V_{fb}) shift with frequency in the depletion region and dispersion of capacitance with frequency in the accumulation region. These metrics, evaluated for all samples based on the C–V characteristics given in Figure 6.24, are presented in Table 6.8. It can be seen that the stretch-out of the C–V response for the 1% treated capacitor is considerably improved, by a factor of three, in comparison to the control sample. The sharper transition from depletion to accumulation is indicative of the improved interface quality resulting from the 1% treatment. However, with increasing $(NH_4)_2S$ concentration, the stretch-out begins to deteriorate. While the 5% treatment presents with a small reduction compared with the 1% treatment, the stretch-out of the 10% treatment approaches that of the control sample. The worst stretch-out characteristic is observed in the C–V response of the 22% treated capacitor.

As with stretch-out, the 1% treated sample presented with the smallest V_{fb} shift of 172 mV. In comparison to the control sample, a factor of two improvement in V_{fb} shift

Table 6.8: Comparison of stretch-out, flatband voltage shift and frequency dispersion in accumulation between Au/Ni/Al_2O_3/p-GaSb MOS capacitors without any treatment (control) and with 1%, 5%, 10% and 22% $(NH_4)_2S$ treatments.

	Stretch-out ($\times 10^{-7}$ F/cm^2.V)	Flatband voltage shift (mV)	Frequency dispersion in accumulation (%/dec)
Control	-0.81	402.18	Accumulation response not observed
1% $(NH_4)_2S$	-2.28	171.84	1.12
5% $(NH_4)_2S$	-1.85	230.01	1.10
10% $(NH_4)_2S$	-0.83	314.60	1.01
22% $(NH_4)_2S$	-0.13	1239.38	Accumulation response not observed

is obtained for the 1% treatment, indicating a reduction of interface trap states from the valence band to midgap. An increase in V_{fb} shift is observed with increasing sulphide concentration. Nonetheless, the values of 230 mV and 315 mV, for the 5% and 10% treated samples respectively, were still better than the control sample. On the other hand, a significant V_{fb} shift of 1240 mV is noted for the 22% treatment.

While the metrics of stretch-out and V_{fb} shift were found to degrade with increasing sulphide concentration, no discernible difference was noted for frequency dispersion in accumulation between the samples. However a closer inspection of the C–V characteristics reveals no apparent accumulation response for the control and 22% treated samples. This is understood from the C–V response at $V_{\mathrm{g}} = -2$ V not presenting with saturation-like capacitance behaviour, one signature of an accumulation response. The fact that this signature is observed for the 1%, 5% and 10% treatments attests to the positive impact of the sulphide treatment at these concentrations. Series resistance due to contacts, substrates and cabling is a common cause of frequency dispersion in accumulation [16]. However, this was unlikely to be the reason since dispersion apart from appearing at the higher frequencies also persisted at lower frequencies where series resistance has no impact [103]. Instead, such frequency dispersion arises from carriers tunnelling into defects/traps with long time constants, which are spatially located in Al_2O_3 dielectric near to the interface, known as border traps [102, 103]. The small dispersion values (see Table 6.8) is one indicator of reduced interface trap density towards the valence band [209].

Cmax Inconsistency

One apparent difference concerns C_{max} in the accumulation region between the chemically treated capacitors. While C_{max} values appear to be similar between the 1% and 5% treated samples at $V_g = -2$ V, the capacitance of the 10% treated sample is 13% lower in

comparison. The 22% treated sample shows a drastic reduction in $C_{\max}$ by approximately 50%. In order to elucidate these differences, physical characterisation in the form of TEM was performed on selected Al_2O_3/GaSb samples. Figure 6.26 depicts the cross-sectional TEM micrographs of the Al_2O_3/p-GaSb structures treated with 1% and 22% sulphide solutions. A number of comparisons and conclusions can be derived from the micrographs with respect to the gate dielectric thickness as well as IL between Al_2O_3 and GaSb. An amorphous Al_2O_3 film with a thickness of 8±0.2 nm, close to the nominal thickness, is observed for the 1% treated sample in Figure 6.26(a). A clear transition from the underlying crystalline GaSb to the amorphous Al_2O_3 film, with no distinct IL, is also evident. In addition, the Al_2O_3 film thickness appears to be uniform across the sample.

The 22% treated sample presents with a very distinct, amorphous IL as illustrated in Figure 6.26(b). This IL, of sizeable thickness, is a result of the more pronounced reaction that occurs between GaSb and $(NH_4)_2S$ solution of higher concentration. It is likely the IL is composed of GaS, a by-product of the sulphide treatment [278]. Voids ranging from $15 - 25$ nm in diameter also appeared to form at non-specific regions along the IL. These voids are believed to be rich in antimony, resulting from the reaction of the sulphide with GaSb [278]. As opposed to highly uniform IL/n-GaSb interfaces resulting from uniform etch characteristics of 22% sulphide in [278], in this work non-uniform etching is seen to occur, as evidenced in Figure 6.26(b), resulting in an etch morphology that is inhomogeneous. It is possible the shorter passivation time of 5 min used in [278] results in a less pronounced, and therefore a more uniform etch, of GaSb. Also, the etch characteristics could be vastly different on n- and p-GaSb. The distinct layers of Al_2O_3,

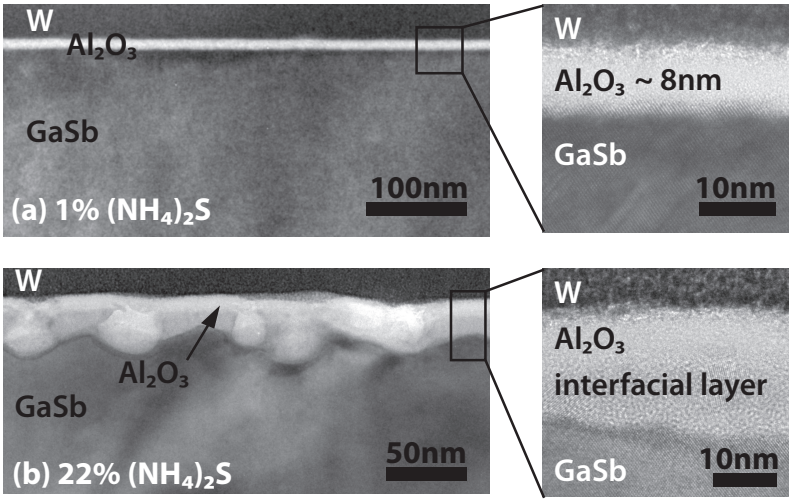


Figure 6.26: Cross-sectional TEM micrographs of (a) 1% and (b) 22% $(NH_4)_2S$ treated Al_2O_3/p-GaSb MOS structures.

IL and GaSb can still be identified from Figure 6.26(b). Moreover, the conformal nature of ALD growth is evidenced from the Al_2O_3 layer following the surface topographical variation resulting from the treatment, leading to an Al_2O_3/GaSb interface that is not flat.

In relation to the C–V characteristics of the 22% sulphide treated sample shown in Figure 6.24(e), the observed lower maximum accumulation capacitance can be attributed to the presence of the IL. The additional interfacial layer capacitance (C_{il}) causes a reduction of the overall capacitance according to:

$$C_{\mathrm{tot}} = \left(\frac{1}{C_{\mathrm{ox}}} + \frac{1}{C_{\mathrm{il}}} + \frac{1}{C_{\mathrm{s}} + C_{\mathrm{it}}} \right)^{-1}, \tag{6.6}$$

where C_{ox} is the gate dielectric capacitance, C_{s} is the semiconductor capacitance and C_{it} is the interface trap capacitance. The presence of an IL also explains the drop in maximum accumulation capacitance of 10% sulphide treated MOS capacitor (see Figure 6.24(d)), although the capacitance is higher than that of the 22% sulphide treated sample. This implies the IL is thinner for the 10% treatment as compared to the 22% treatment. Therefore, the IL thickness appears to be dependent on the concentration of the sulphide solution, with thicker IL resulting from solutions of higher concentrations. Papis *et al.* [279] reached a similar conclusion and further found the thickness of IL to be dependent on the passivation duration (a parameter not investigated in this work). The capacitance loss, arising from the presence of such interfacial layers, would compromise gate control in p-channel flatband MOSFETs resulting in ineffective transistor operation.

TEM imaging was also undertaken on an untreated Al_2O_3/GaSb sample to identify the presence of native oxides. As observed from Figure 6.27 no distinguishable IL was present, similar to the interface of the 1% treated sample. It is possible the native oxides inherent to GaSb have been removed as a result of the ALD "self-cleaning" effect [285, 286]. However, based on the TEM alone it cannot be concluded whether the GaSb surface is

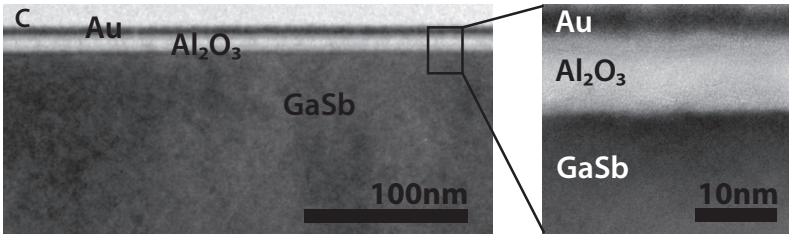


Figure 6.27: Cross-sectional TEM micrograph of untreated (native surface) Al_2O_3/p-GaSb MOS structure.

free of native oxides, given the observed C–V response (see Figure 6.24(a)). What is clear though is a higher interface roughness for the untreated sample in comparison to the 1% treated sample.

Gate Leakage

The aforementioned TEM analysis can also be correlated with the leakage characteristics of the capacitors, plotted in Figure 6.28. At $V_g = -1\,V$, the untreated as well as the 1% and 22% treated samples exhibited gate leakage current density (J_g) in the range of $1-7\times10^{-8}\,A/cm^2$. On the other hand J_g increased to $\sim2\times10^{-6}\,A/cm^2$ for the 5% and 10% treated samples. The good leakage characteristics of the untreated and 1% treated samples is to be expected given the good interface quality observed from TEM. Interestingly, despite the poor interface of the 22% treated sample, plagued by non-uniformities and the presence of an IL, the leakage characteristics were comparable to those of the good interfaces. It is possible the substantially thick IL is able to compensate for the poor interface formation. Although the 10% treated sample is also likely to have an IL as explained earlier, this is significantly thinner given the drop in C_{max} was not as significant as that of the 22% treated sample. This accounts for the increased leakage. The 5% treatment, despite presenting with a C–V response comparable to the 1% treatment, appears to be more affected by gate leakage. Nonetheless, all leakage values were still found to be orders of magnitude lower than the J_g requirement of $1.25\times10^{3}\,A/cm^2$ set out in the International Technology Roadmap for Semiconductors (ITRS) [287], indicating the high quality of ALD grown Al_2O_3 dielectric.

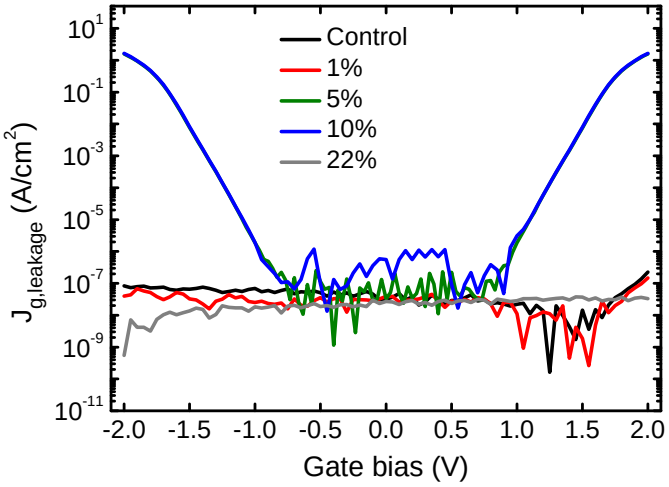


Figure 6.28: I–V characteristics of control (untreated) and 1%, 5%, 10% and 22% $(NH_4)_2S$ treated Al_2O_3/p-GaSb MOS capacitors.

Measured vs. Theoretical C-V

Revisiting the C–V response in accumulation, a measure of the effectiveness of the $(NH_4)_2S$ treatment is to compare the extent of deviation between the measured and theoretical capacitance in accumulation at $V_g = -2\,V$. As before, the theoretical C_{max} in accumulation was determined from a 1D-PS C–V simulation. For both simulation-based and simple analytical calculations, a physical Al_2O_3 thickness of 8 nm (based on TEM) was used. The dielectric constants of Al_2O_3 and SiO_2 were taken to be 8 and 3.9, respectively [26]. Based on the 1D-PS simulation shown in Figure 6.24(f), the accumulation capacitance at $V_g = -2\,V$ was determined to be $0.822\,\mu F/cm^2$, with a corresponding CET of 4.2 nm. A C_{ox} of $0.885\,\mu F/cm^2$, as indicated in Figure 6.24(f), was theoretically determined for Al_2O_3 dielectric, giving a EOT of 3.9 nm.

In theory C_{max} in accumulation should approximate to C_{ox}. This holds if the accumulation/inversion layer charge is assumed to be a sheet charge of zero thickness at the dielectric interface, for which $C_s \gg C_{ox}$. However, in real MOS systems this assumption is invalid. The adoption of high-k dielectrics with t_{ox} in the nanometre range results in an "electrical" oxide thickness (t_{ox}/ε_{ox}) that is comparable to effective "electrical" accumulation/inversion layer thickness (t_s/ε_s). Equally important are the effects of DOS and charge quantisation in semiconductors [99], as discussed in Section 2.3.4.1. A smaller DOS in the conduction or valence band would act to reduce the density of states capacitance, otherwise known as quantum capacitance. The effect of charge quantisation is to move the charge centroid of the accumulation/inversion layer further into the semiconductor, away from the dielectric interface, and in the process decrease the centroid capacitance. For a lower DOS the charge centroid moves further away from the interface. Taking the contributions from both these effects into account results in a CET that is larger than the EOT. This forms the basis of C_{max} in accumulation being reduced from the oxide capacitance as given in Equation (2.39). For the purpose of comparison, the difference between CET and EOT will be referred to as equivalent oxide thickness correction (EOT_{corr}). The EOT_{corr} for the Al_2O_3/p-GaSb system is 0.3 nm. This is much smaller than the EOT_{corr} for devices on n-$In_{0.53}Ga_{0.47}As$, which lies between 1.1 nm and 1.5 nm [26]. The smaller correction for GaSb is due to the high valence band DOS ($N_V^{DOS} = 1.8\times10^{19}\,cm^{-3}$) which is two orders of magnitude greater than the conduction band DOS in $In_{0.53}Ga_{0.47}As$ [68].

It is noted the measured C_{max} values of $0.845\,\mu F/cm^2$ and $0.839\,\mu F/cm^2$ for the 1% and 5% treated samples are higher compared to the theoretical C_{max} of $0.822\,\mu F/cm^2$ derived using 1D-PS simulation (see Figure 6.24(f)). This difference is accounted for by the presence of trap states close to the valence band in GaSb [26, 99]. The deviation of the

measured to simulated C_{max} is $0.023\,\mu\mathrm{F/cm^2}$ and $0.017\,\mu\mathrm{F/cm^2}$ for the 1% and 5% treated samples. This points to the latter sample having a lower density of traps at the valence band edge. As for the other samples, the deviation in capacitance was not evaluated for the following reasons. The untreated sample is likely not to enter accumulation as discussed earlier, and therefore invalidates the use of this metric. In the case of the 10% and 22% treated samples, the fact that the measured C_{max} in accumulation is smaller than the simulated capacitance holds no physical meaning. To enable an accurate determination of the deviation, the simulation needs to take into account the capacitive contribution of the IL presence in these samples.

Based on theoretical modelling, a feature that is observed for III-Vs but not Si is an asymmetrical LF C–V response [26, 99, 282]. This asymmetry is attributed to the differing DOS in the conduction and valence bands. In addition DOS in the conduction band tends to be typically lower, by as much as two orders of magnitude, than the valence band [68]. This originates from the carrier effective mass being considerably lower for electrons in contrast to holes. It is this lower electron effective mass, and hence higher electron mobility, that has placed III-V as a frontrunner to succeed Si as the n-channel solution in digital logic. Owing to DOS being lower on the conduction band side, the capacitance for electron inversion/accumulation layers should be lower in comparison to the hole inversion/accumulation layers, at least in theory. In Si the DOS are comparable between the valence and conduction band [68], and therefore, does not present with this asymmetry. In line with the theoretical considerations, the C–V responses of 1% and 5% treated samples are indeed asymmetrical. This asymmetry, quantified as the difference in capacitance between gate biases of $-2\,\mathrm{V}$ and $+2\,\mathrm{V}$, is $0.013\,\mu\mathrm{F/cm^2}$ and $0.007\,\mu\mathrm{F/cm^2}$ for the 1% and 5% treated samples respectively. Given the valence and conduction band DOS differ by about two orders of magnitude, these asymmetry numbers appear to be rather small. The capacitive contribution of interface states, energetically aligned with the conduction band edge to C_{inv}, can be drawn on to account for this discrepancy.

Temperature Dependence

Although C_{max} in accumulation appears to be reaching C_{ox} for the 1% and 5% treated samples (see Figure 6.24), it does not necessarily mean an accumulation layer is formed. This could also be consistent with an interface trap dominated capacitance response. For an interface dominated by a high trap density, accumulation of free carriers is small. As a result C_{s} is small in comparison to C_{it}. However, at low frequencies C_{max} in accumulation approaches the value of C_{ox} due to the high trap capacitance, regardless of the small free carrier response [36, 288]. One experimental method to determine if C_{max} in accumulation

originates from free carrier response, or rather a consequence of a large C_{it}, is to perform a temperature-dependent admittance study [36]. Such an investigation was performed on the 1% $(NH_4)_2S$-treated sample and the results of the 1 MHz C–V response as a function of temperature (-40°C to 20°C) is depicted in Figure 6.29. The capacitance dispersion with temperature at $V_g = -1.5$ V was found to be small (∼3.9%). This implies the observed accumulation behaviour is most likely a result of free carrier accumulation as opposed to trap-induced response. The same can be said of the 5% treated sample. In addition, the flatband voltage shift with temperature was determined to be 128 mV. The minimal vertical and horizontal shifts of the C–V curves with temperature, in accumulation and depletion respectively, are characteristic of a low trap density below midgap. It further suggests the Fermi level at the Al_2O_3/GaSb interface can be swept in the lower part of E_g to the valence band edge.

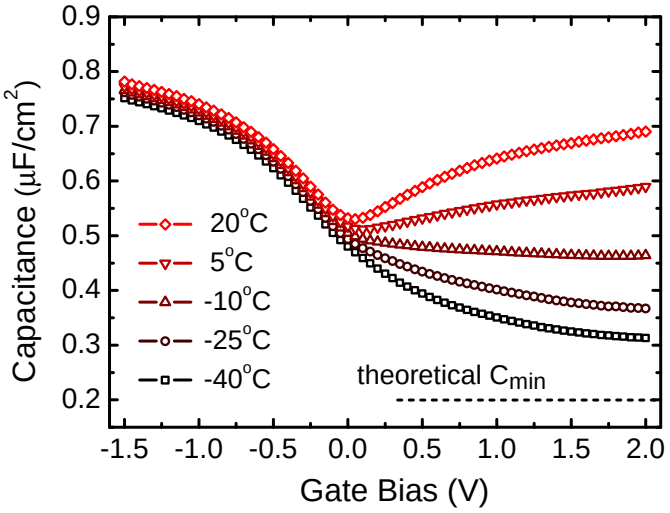


Figure 6.29: 1 MHz C–V response of Au/Ni/Al_2O_3/p-GaSb MOS capacitor treated with 1% $(NH_4)_2S$ as a function of temperature (-40°C to 20°C).

The capacitance dispersion observed in the gate bias range of +0.5 V to +2 V in Figure 6.29 is characteristic of interface traps. The reduction in capacitance with temperature in this voltage range is related to the emission/capture rate of trap states, τ_{it}, which has an exponential dependence on 1/T (see Equation (2.41)). Therefore, as the temperature of the sample is lowered, the interface trap response becomes suppressed in the upper half of E_g towards the conduction band. As a result of the reduced contribution from C_{it}, the inversion capacitance drops and the C–V response approximates towards a HF behaviour

at the lower temperatures. This is observed in Figure 6.29. Furthermore, unlike the pinned ALD-Al_2O_3/p-GaSb interface in the temperature range of 200–300 K in [259], the Fermi level of the 1% treated sample shows clear movement from accumulation to inversion with decreasing temperature. However, the theoretical minimum C_{inv} of 0.203 μF/cm^2 (1D-PS simulation) is still not reached at $V_{\text{g}} = +2$ V, despite the 60 K temperature reduction from RT. This suggests the trap response has not been completely frozen out at -40°C and thus still manifests as a capacitive contribution, albeit much reduced. These results are further evidence that the 1 MHz response at RT is one of false inversion. However, the dominance of traps in the upper half of E_{g} is more pronounced for the 5% compared to the 1% treated sample. This is inferred from the lower capacitance of the 1 MHz curve in the bias range of 0 V to +2 V (see Figure 6.24), which further underscores the efficacy of the 1% treatment for passivation of trap states.

Interface Trap Density

To quantify the trap distribution at the interface of Al_2O_3/GaSb treated with 1% $(NH_4)_2S$ a temperature-modified version [26] of the high-low frequency C–V method [97], discussed in Section 4.4.1.3, is employed. In the method the 1 kHz C–V response at RT was taken as C_{LF}. In contrast the high frequency capacitance measurement (C_{HF}) was based on the 1 MHz C–V response obtained at -40°C in this work, to minimise the interface trap response, thereby enabling a more accurate D_{it} determination. D_{it} was then extracted using Equation (4.20) over a limited bias range (-1.5 V to 0 V), since the method is only valid from weak inversion towards the majority carrier band edge [97]. The band bending as function of gate bias, ψ_{s}–V_{g}, was obtained from the Berglund integral (see Equation (4.21)), which then enabled the trap energy position at the Al_2O_3/GaSb interface to be calculated using Equation (4.22). The resulting D_{it} distribution of the 1% treated sample is shown in Figure 6.30. A U-shaped profile is observed with a minimum D_{it} of 4×10^{12} cm^{-2}eV^{-1} occurring at $E_{\text{v}} + 0.27$ eV. Such a D_{it} profile close to the midgap is suggestive of a lower SS and therefore improved off-state performance of inversion-type p-channel MOSFETs [84].

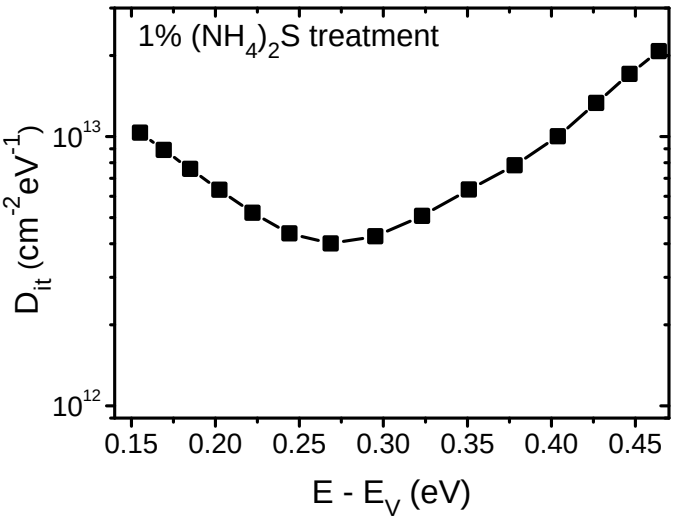


Figure 6.30: Interface trap density (D_{it}) distribution from weak inversion towards the majority carrier band edge of the 1% $(NH_4)_2S$) treated MOS capacitor determined from the temperature modified high-low frequency method.

6.5.3.2 Surface Topography Study

Ex-situ AFM scans were carried out on samples subjected to the variety of chemical treatments to determine the impact of sample processing following only sulphide treatment.

Surface Roughness

Figure 6.31 illustrates the RMS roughness results determined from AFM images of $5\,\mu$m$\times 5\,\mu$m scan area from individual samples. An untreated sample served as a control during the study, for comparing the changes in roughness prior to and after treatment. It is observed the resulting surface roughness of the 1% and 5% treated samples for all immersion times is minimally changed from the control sample and lie in the range of 1.37–1.45 nm. However, at the higher concentrations of 10% and 22% roughness presents with a monotonic increase with immersion times. For the sample treated in 22% $(NH_4)_2S$ for 10 min, a large roughness of 3.25 nm results. Although the trend appears to be somewhat similar to sulphide treated InSb of [264], the roughness of chemically treated GaSb appeared to be more pronounced. Even the untreated GaSb sample had a roughness that is $\sim 3\times$ higher than the untreated InSb sample in [264]. While on the one hand GaSb appears to be more suspectable to native oxide formation, on the other $(NH_4)_2S$ seems to present with increased chemical activity on GaSb. The roughness trend of GaSb is also coincident with the interface roughness deduced from the TEM analysis earlier. The increase in roughness with increasing sulphide concentration at a fixed sample immersion time of 10 min is in support of the C–V responses shown in Figure 6.24.

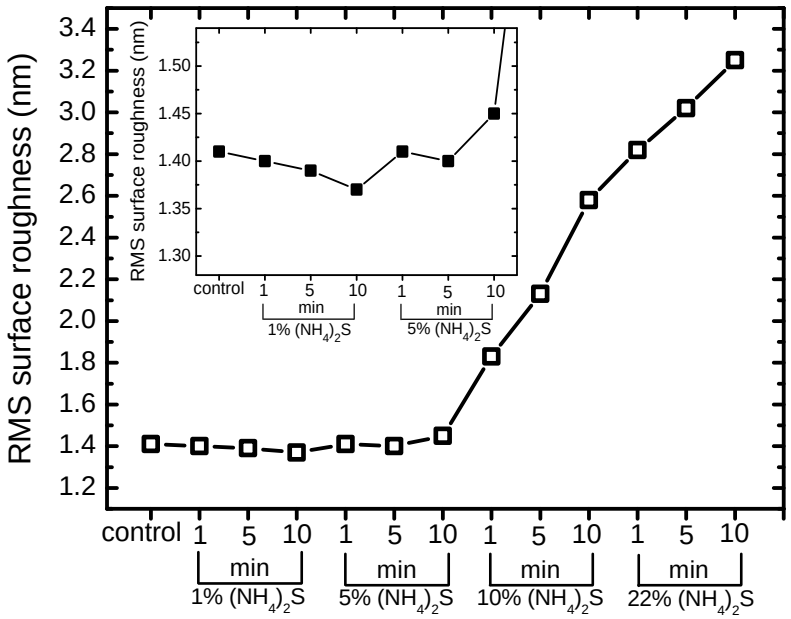


Figure 6.31: Root-mean-square (RMS) roughness of p-GaSb (7-144) samples subjected to various $(NH_4)_2S$ treatments based on AFM measurements. The inset depicts a clearer picture of the roughness variation in the treatment range between control and 5% $(NH_4)_2S$.

Etch Characteristics

Apart from the surface roughness, the etch characteristics of GaSb is also an important consideration for the realisation of GaSb transistors. The thickness of GaSb etched as a function of the chemical treatments is listed in Table 6.9, based on AFM step height measurements. The etched layer thickness was generally observed to increase with increasing immersion times and $(NH_4)_2S$ concentration. For samples (i) to (x) the thickness of the etched layer showed a gradual increase with both concentration and immersion times. In contrast, a sharp increase in the etch rate was noted for samples (xi) and (xii) treated with 22% $(NH_4)_2S$ for 5 and 10 mins respectively. An etched layer thickness of approximately 71 nm was obtained for sample (xii). For the optimum sample treatment of 1% sulphide for 10 mins, an etched layer thickness of approximately 5 nm was measured. Since this was the treatment to be employed in the InGaSb-channel device wafers of this work, it was important to account for this thickness during the MBE growth of the GaSb cap.

Table 6.9: Etch depth of p-GaSb (7-144) samples subjected to various $(NH_4)_2S$ treatments based on AFM measurements.

Sample number	Concentration (% $(NH_4)_2S$ in H_2O)	Immersion time (min)	Etch depth (nm)
(i)	1	1	0.55
(ii)	1	5	1.12
(iii)	1	10	5.22
(iv)	5	1	1.02
(v)	5	5	2.53
(vi)	5	10	9.2
(vii)	10	1	5.11
(viii)	10	5	6.4
(ix)	10	10	12.37
(x)	22	1	13.14
(xi)	22	5	40
(xii)	22	10	71.23

6.5.4 Summary

Interface passivation technologies are crucial to achieving high quality interfaces in high-k/III-V MOS systems. This is a key requirement of low-power, high performance GaSb MOSFETs to enable high $I_{\mathrm{d,sat}}$ and small SS.

Ex-situ aqueous $(NH_4)_2S$ treatments were employed for Al_2O_3/GaSb MOS structure. The conditions of 1%, 5%, 10% and 22% diluted sulphide solutions, for a fixed sample immersion time of 10 min, were investigated. An untreated sample served as the control in the study. Based on admittance measurements and TEM analysis the 1% $(NH_4)_2S$ treatment was determined to be optimum for implementation in devices. The effect of interface traps and its distribution in the relevant parts of the bandgap were evaluated from RT and temperature-dependent admittance data. Samples treated at the higher sulphide concentrations suffered from the formation of an IL and increased interface roughness. The combination of these two effects degraded the interface quality, which was reflected in the admittance characteristics.

Surface topography examination, by means of AFM, revealed the roughness character of GaSb subject to a multitude of treatments with both the concentration and immersion time varied. The thickness of the etched GaSb layers resulting from these chemical

treatments were further quantified. This enables the treatment-dependent etch to be accounted for during growth of GaSb-terminated III-V epi-layers based on the chosen treatment. For the GaSb-based transistors of this work, the 1% $(NH_4)_2S$ treatment for 10 min was employed. This meant the GaSb cap needs to be at least 8 nm to allow for a 2 nm cap to be still present following chemical treatment and was taken into account during the design of device layers.

6.6 Chapter Summary

This chapter detailed the key process modules that were developed and optimised for integration into a device flow for transistor realisation. Three modules were developed and investigated for S/D and gate stack engineering; ohmic contacts, gate-to-source/drain gap, L_{side}, scaling and high-k/GaSb interface passivation. As a summary of each developed process module was included at the end of each section, these will not be repeated here. Rather, this chapter summary is intended to highlight the process conditions of each module as implemented in the realisation of devices based around InGaAs- and InGaSb-channels, discussed in the following chapter. Ohmic contacts to InGaAs-channel devices was based on Au(5 nm)/Zn(40 nm)/Au(150 nm) metallisation annealed at 380°C for 120 s, with a de-oxidation condition of 1 HCl : 4 H_2O, 30 s. For contacts to InGaSb-channel devices the de-oxidation procedure of 1 HCl : 2 H_2O for 60 s is employed prior to Au(5 nm)/Zn(40 nm)/Au(100 nm) metallisation annealed at 340°C for 120s. The lithographically-aligned double ohmic patterning fabrication process was applied to selected InGaAs-channel devices to realise L_{side} gaps of 70 nm, 200 nm and 465 nm. The gate stack on InGaSb-channel devices was obtained by treating the surface of the GaSb capping layer in 1% $(NH_4)_2S$ solution for 10 min at 295 K prior to ALD of Al_2O_3.

7

Device Results

7.1 Introduction

The demonstration of a functional transistor serves as the ultimate test in evaluating semiconductor material systems for MOS applications. In accordance p-MOSFETs, based around III-V compound semiconductors and featuring high-k/metal gate (HKMG) stacks, are demonstrated in this work. Both InGaAs and InGaSb are investigated as channel materials. The former is a natural choice for investigation, since InGaAs is primed to be the first III-V entry into CMOS as the n-channel solution. The perpetual need for faster switching performance in digital logic also motivates an examination of III-Vs with higher mobilities. Antimonides, such as GaSb and InSb, as well as the combination of the two binaries in different stoichiometries are attractive as they have the highest electron/hole mobilities amongst all III-V materials [68].

The design of MOSFET layers, employing the flatband device architecture [224], are based on the pre-MOSFET structures investigated in Chapter 5. All devices employ the buried-channel design, in which wider bandgap cap/barrier layers are introduced between the gate dielectric and the channel to mitigate the interface and dielectric-related mobility degradation [81, 92, 180, 181]. The tradeoff with the buried-channel structure is the reduction in the gate modulation efficiency of the channel charge. In order to preserve gate control, the wider bandgap cap/barrier layers are kept intentionally thin in the designs.

Within the buried-channel design two aspects of layer optimisation - strain and doping strategy - are subject to study. The majority of III-Vs have hole mobilities that are comparable to Si, with the exception of the antimonides which show 2–3$\times$ higher mobility [68]. Therefore, without the addition of strain for enhanced channel mobility it is difficult

to envisage III-Vs having a competitive advantage over Si from a pMOS point of view. For this reason biaxial compressively strained channels are adopted in the designs, with variations in strain explored. In addition the site of the δ-doping plane, whether above (non-inverted) or below (inverted) the channel, was found to affect the resulting carrier transport in pre-MOSFETs otherwise similarly designed. Non-inverted pre-MOSFETs exhibited a higher μ_{h} and carrier density than the inverted design, by 37% and 20% respectively (see Section 5.4). This translates to lower R_{sd}, and higher $I_{\mathrm{d,sat}}$ and g_{m} in devices. The additional spacer thickness of the non-inverted design however would add to CET, resulting in gate control degradation. In view of this both inverted and non-inverted designs are examined.

As highlighted in Section 1.2, achieving a high quality gate stack with low D_{it} remains a major obstacle to III-V integration in CMOS. A high trap distribution in the relevant parts of the bandgap swept by the Fermi level during transistor operation negatively impacts the on and off-state performance metrics [223, 289, 290]. To enable a high quality interface with minimum trap density, three components of the gate stack need attention. The first is the III-V surface on which the dielectric is deposited. As pointed out in Section 5.2, Al-containing surface layers are undesirable due to their higher susceptibility to oxidation [206, 207] and thus are avoided in the designs. In addition, the choice of dielectric has a strong bearing on the resulting high-k/III-V interface. Trivalent dielectrics (e.g. Al_2O_3) have been reported as being more effective passivants of III-V surfaces as opposed to tetravalent dielectrics (e.g. hafnium oxide (HfO_2)) [35, 38]. This is a direct result of the lower subcutaneous oxidation [38] and smaller roughness [23] at the Al_2O_3/III-V interface. Moreover, the CBO/VBO of Al_2O_3 on $In_{0.53}Ga_{0.47}As$ and GaSb are sufficient to minimise gate leakage [85, 257, 291, 292]. Al_2O_3 is therefore adopted as the gate dielectric in all devices. The final piece of the gate stack puzzle is interface passivation. For InGaAs-channel devices the technique of arsenic capping [29] is employed. On the other hand, sulphide treatment is applied to InGaSb-channel MOSFETs based on the $(NH_4)_2S$ interface passivation study presented previously (see Section 6.5). Details of the passivation conditions will be discussed in the relevant sections of this chapter.

In this work, III-V epi-layers were grown via MBE, while ALD was used for Al_2O_3 gate dielectric deposition*. Transistor fabrication† was based on the device layouts and

*For the InGaAs-channel device wafers, the III-V epi was grown at the University of Glasgow and the gate dielectric was deposited at Stanford University. For the InGaSb-channel device wafers, the epi was obtained from Texas State University, while dielectric deposition was performed at Tyndall National Institute.

†Device fabrication, and subsequent electrical characterisation, were undertaken at the University of Glasgow.

accompanying process flows presented in Chapter 6, incorporating the developed process modules where relevant. Different aspects of device characteristics were appraised. The long-channel behaviour of all devices were evaluated to exclude SCE on the performance metrics. This also serves to establish a baseline comparison between devices of different III-V epi and layer designs. In addition the impact of geometrical scaling on transistor metrics was assessed on selected devices. Figures of merit are evaluated based on the analysis techniques covered in Chapter 4. In the following sections the performance of InGaAs and InGaSb-channel devices is presented and discussed.

7.2 InGaAs-Channel Device Performance

7.2.1 Device Specification and Processing

Figure 7.1 illustrates the layer structure of the InGaAs-channel device wafers. Both wafers feature a 10 nm thick, 75% In-content channel (1.4% strain). In line with the buried-channel design a 2 nm $In_{0.53}Ga_{0.47}As$ cap and a 2 nm $In_{0.52}Al_{0.48}As$ barrier layer are incorporated in the device layers. The Al-free cap layer is intended to serve as a passivation layer on InAlAs, suppressing aluminium oxidation and providing a much improved

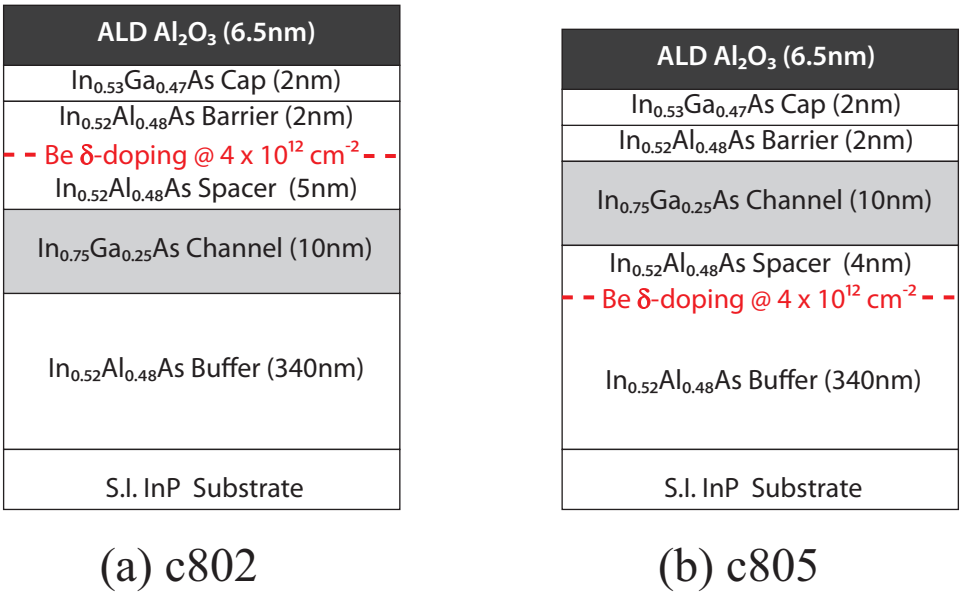


Figure 7.1: Layer structure of InGaAs-channel device wafers based on the implant-free, flatband MOSFET architecture featuring a 10 nm thick $In_{0.75}Ga_{0.25}As$ channel under a 1.4% biaxial compressive strain: (a) c802 the non-inverted design and (b) c805 the inverted design.

interface with the high-k dielectric [206]. Device wafer c802 was based on the non-inverted design whereas the inverted design concept was applied in the c805 device wafer.

The InGaAs-channel wafers were capped in-situ with a 80 nm amorphous arsenic (As_2) layer immediately after epi-layer growth, to protect the surface from contamination and oxidation resulting from air-exposure during the transfer of wafers between the MBE tool and the ALD chamber [29]. Prior to the gate dielectric deposition, an in-situ thermal desorption (decapping) procedure was employed to remove the As_2 capping layer in the ALD chamber [293]. The decapping process was carried out under high vacuum and with the flow of forming gas (5% H_2/ 95% N_2) at a pressure of 1 Torr [29]. In order to remove any residual moisture present in the As_2 layer the capped wafers were initially baked for $\geq$ 1 hr at 150°C, and then gradually heated up at a rate of 5–6 °C/min in the ALD tool. The complete removal of the As_2 layer occurred at approximately 370°C. Two experimental observations were used for the end-point detection of the decapping procedure [293], the first of which is a change in the background pressure of the ALD chamber. During the desorption process the pressure goes through a sharp rise followed by a sudden fall back to the initial chamber pressure of 2×10^{-7} Torr, with the peak of the pressure pulse coinciding with the decapping temperature. In addition, the chemical binding energy of the As $3d$ peaks before and after the decapping process was measured using in-situ XPS. The shift in the binding energy, by approximately -0.8 eV, confirmed the completion of the As_2 decapping. Following decapping, Al_2O_3 films with a nominal thickness of 6 nm were deposited by 60 cycles of ALD at 270°C using TMA and H_2O precursors, with a TMA-first pulsing during the growth process [294]. TEM measurements were conducted and the physical oxide thickness was found to be 6.5 nm [173].

A common fabrication process comprising a combined gate/marker level and an ohmic level in a GF device flow (see Figure 6.3) was employed on both the c802 and c805 wafers for transistor realisation. Details of the processing steps can be found in Appendix A.2. The resulting devices featured several gate lengths ranging from 500 nm up to 20 μm, at a fixed L_{side} of 1 μm and gate width (W_g) of 100 μm. Scaled L_{side} devices were fabricated on the c805 wafer for proof of concept. Transistor fabrication consisted chronologically of a combined gate/marker level, stripe ohmic level and a large ohmic level (see Figure 6.22). Comprehensive details of this fabrication flow can be found in Appendix A.3. Devices featuring several L_{side} gaps - 70 nm, 200 nm, 465 nm and 1 μm - were fabricated for investigation. The effect of differing transistor footprints on the resulting device characteristics are discussed in the following sections.

7.2.2 c802 Device Results and Discussion

Illustrated in Figure 7.2 are the DC electrical characteristics of a long-channel transistor, with $L_g = 2\,\mu$m and $L_{side} = 1\,\mu$m, fabricated on the c802 wafer. Observations based on first inspection are initially presented, following which device characteristics are analysed in more detail.

Figure 7.2(a) shows the output characteristics of the device. A well-behaved I–V response with good saturation and switch-off characteristics is demonstrated. The device operates in depletion-mode and has a threshold voltage (V_{th}) of $+0.189\,$V determined from the linear extrapolation of the I_d–V_g characteristic at $V_d = -50\,$mV. The transistor drive currents however are rather modest. A maximum $I_{d,sat}$ of $1.14\,$mA/mm was obtained at $V_g = -3\,$V and $V_d = -2\,$V. It is also evident that the device suffers from a relatively large source/drain series resistance, noticeable from the slope of the curves in the linear region of the I_d–V_d plot. Consequently, a high R_{on} of $845\,\text{k}\Omega.\mu$m was extracted from the plot at a gate bias of $-3\,$V and a drain bias of $-50\,$mV.

The I_d–V_g curve is plotted on a semi-logarithmic scale in Figure 7.2(b) to highlight the subthreshold characteristics. The device presents with a minimum SS of $301\,$mV/dec at $V_d = -50\,$mV. Similar swings are maintained up to a drain bias of $-0.4\,$V, beyond which SS degrades with increasing drain bias. A factor of two increase in the swing ($SS \sim 601\,$mV/dec) is noted at a drain bias of $-2\,$V.

In the context of this work, I_{on}/I_{off} is defined as the ratio between the drain current at the gate biases of $V_{th} + 2/3\,V_d$ (on-state) and $V_{th} - 1/3\,V_d$ (off-state) at a specified drain bias [295]. Based on this, the drain current ratio between the on-state and the off-state (I_{on}/I_{off}) is determined to be approximately 74 at $V_d = -2\,$V. This poor showing is in part attributed to the subthreshold performance, which presents with high SS and off-state leakage clearly visible in Figure 7.2(b), and in part to the performance-limited on-state. A notable feature of the device characteristics is the appearance of multiple transconductance peaks at all drain biases, as observed in Figure 7.2(c). In the following sections the various aspects of device performance are discussed.

7.2.2.1 Threshold Voltage

A key requirement of digital logic is E-mode operation [296], which translates to a negative threshold voltage in p-MOSFETs. Contrary to this, devices on c802 wafer present with a positive V_{th}, which could be a result of the device electrostatics originating from the layer design. This was investigated through one-dimensional self-consistent Poisson-Schrödinger

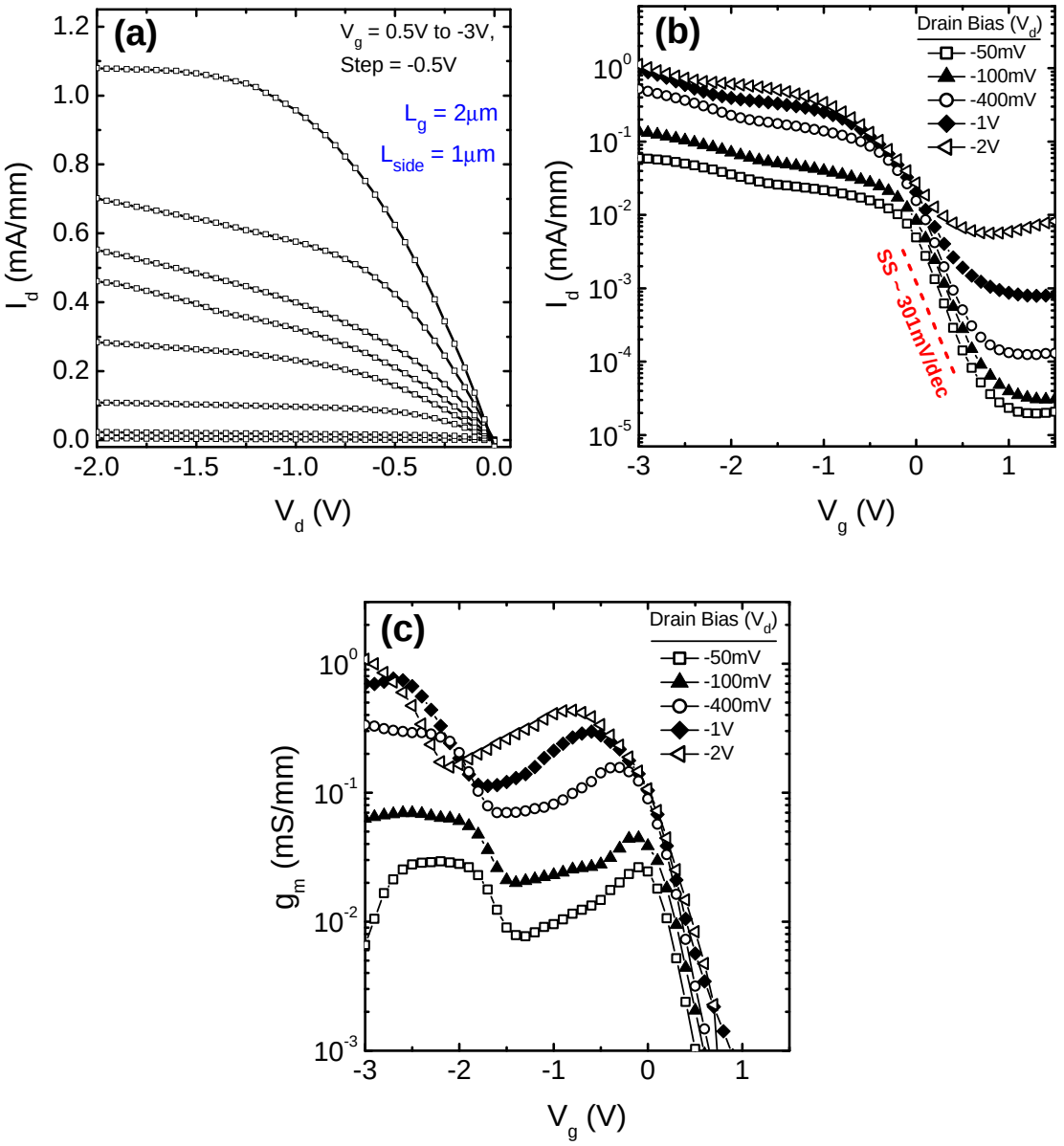


Figure 7.2: Measured (a) output - I_{d}–V_{d}, (b) subthreshold - $\log(I_{\mathrm{d}})$–V_{g} and (c) transfer - $\log(g_{\mathrm{m}})$–V_{g} characteristics of a 2 μm - L_{g}, 1 μm - L_{side} p-MOSFET realised on a c802 device wafer. The device figures of merit are as follows (measurement conditions are given in parentheses): $V_{\mathrm{th}} = +0.189\,\mathrm{V}$ (linear extrapolation of the I_{d}–V_{g} curve at $V_{\mathrm{d}} = -50\,\mathrm{mV}$), $I_{\mathrm{d,sat}} = 1.14\,\mathrm{mA/mm}$ ($V_{\mathrm{g}} = -3\,\mathrm{V}$ and $V_{\mathrm{d}} = -2\,\mathrm{V}$), $R_{\mathrm{on}} = 845\,\mathrm{k\Omega}.\mu\mathrm{m}$ ($V_{\mathrm{g}} = -3\,\mathrm{V}$ and $V_{\mathrm{d}} = -50\,\mathrm{mV}$), $SS = 301\,\mathrm{mV/dec}$ ($V_{\mathrm{g}} = 0.2$ to $0.5\,\mathrm{V}$ and $V_{\mathrm{d}} = -50\,\mathrm{mV}$) and $I_{\mathrm{on}}/I_{\mathrm{off}} = 74$ (I_{off} at $V_{\mathrm{g}} = V_{\mathrm{th}} - 1/3V_{\mathrm{d}}$, I_{on} at $V_{\mathrm{g}} = V_{\mathrm{th}} + 2/3\,\mathrm{V_d}$ and $V_{\mathrm{d}} = -2\,\mathrm{V}$).

(1D-PS) simulations[‡] [280], enabling the carrier density in the channel to be calculated as a function of gate bias. From the calculations (see Figure 7.3), the threshold voltage was determined to be −0.38 V. This suggests the layer design in fact favours E-mode operation. The experimental V_{th} however is approximately 0.57 V more positive than the simulation, which implies non-ideal effects are in play. Such non-idealities manifest in the gate stack as interface trap charge (Q_{it}), fixed oxide charge (Q_{fixed}) and fixed metal/dielectric interface charge ($Q_{\mathrm{m/d}}$). The total charge (Q_{tot}) of the MOS stack comprising these individual contributions is given as

$$Q_{\mathrm{tot}} = Q_{\mathrm{it}} + Q_{\mathrm{fixed}} + Q_{\mathrm{m/d}}\,. \tag{7.1}$$

A negative Q_{tot} of sufficient magnitude, arising from the combined effect of these charge contributions, would account for the disparity between the idealised and experimental value of V_{th}. The following discussion serves to elucidate the contributions of the various components given in Equation (7.1) to the observed V_{th} shift.

The impact of Q_{it} on the threshold voltage is dependent on the traps aligned with the Fermi level at the dielectric/semiconductor interface. Relevant to this is the distribution of interface traps across the semiconductor bandgap and the type of interface defect (donor or acceptor), which in turn dictate the magnitude and polarity of the resulting charge contribution [289, 290]. Illustrated in Figure 7.4 is the energy band diagram of a c802

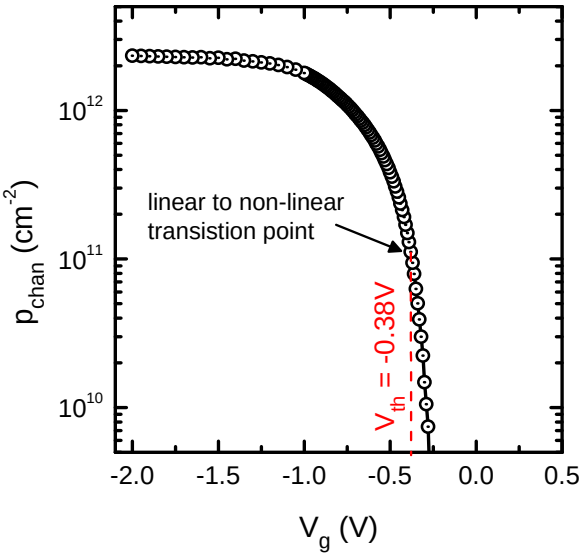


Figure 7.3: Gate bias dependence of channel carrier density calculated from Poisson-Schrödinger simulations for the c802 device wafer. Threshold voltage (V_{th}) is defined as the point at which the log-scale plot changes from a linear to a non-linear trace.

[‡]Band offsets at the $Al_2O_3/In_{0.53}Ga_{0.47}As$ interface were based on [291]. The EWF of aluminium (gate metal) was taken to be 3.9 eV based on [233].

MOSFET biased at threshold ($V_g = V_{th}$). The specific interface trap energy position, E_{tr}, that is addressed at a given bias condition is informed by the Fermi level position at the $Al_2O_3/In_{0.53}Ga_{0.47}As$ interface. The outcome of V_{th} is only determined by traps whose energy positions coincide with the bias condition of $V_g \leq V_{th}$, demarcated in the band diagram. Given the trap energy position at threshold is $-0.34\,eV$ (E_{tr} is defined as zero at the conduction band edge and negative within the bandgap), the traps of interest to V_{th} lie at energies $\geq -0.34\,eV$. For a measure of the trap density, the D_{it} data of [173] is employed as it best reflects the $Al_2O_3/In_{0.53}Ga_{0.47}As$ interface of devices in this work. Part of this is attributed to the III-V epi and gate dielectric grown from identical sources between the devices. Moreover, both interfaces were subject to a common interface passivation technique of arsenic capping/de-capping for which conditions were similar. This therefore rendered the interfaces to be identically passivated. For the purpose of this analysis and for later discussions, the D_{it} distribution from [173] is reproduced in Figure 7.4.

Equally important to the foregoing argument is the nature and occupation of the interface traps swept by the Fermi level during device operation. A positive shift in V_{th} will only arise for a negative charge contribution resulting from filled acceptor states. Experimental investigations of Al_2O_3/ $In_{0.53}Ga_{0.47}As$ MOS systems have identified donor defect states to be present throughout the bandgap, while acceptor states reside inside the conduction band [297, 298]. Under such a donor/acceptor distribution, the resulting trap

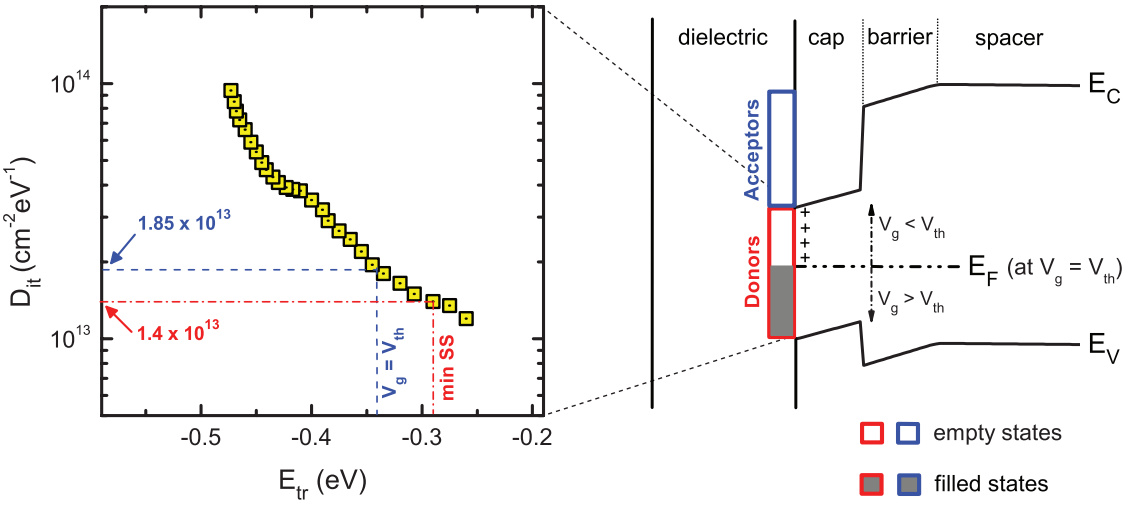


Figure 7.4: Illustration of the energy band diagram of the c802 device at threshold bias ($V_g = V_{th}$) and the $D_{it}(E)$ profile for an $Al_2O_3/In_{0.53}Ga_{0.47}As$ MOS system (adapted from [173]). Also included is the donor/acceptor state distribution at the interface and the associated occupation of traps at threshold bias. E_c, E_v, and E_F are the conduction band and valence band edges and the Fermi level respectively. Trap energy position (E_{tr}) is defined as zero at the conduction band edge and negative within the bandgap.

charge character at the $Al_2O_3/In_{0.53}Ga_{0.47}As$ interface for the bias condition of threshold is indicated in the band diagram of Figure 7.4. The only charge contribution arises from empty donor states above E_{tr}, which results in positively charged interface traps. The 1D-PS simulation from before was repeated to include this charge contribution. For an accurate simulation, the energy dependent trap distribution $D_{it}(E)$ should be taken into account. However, due to limitations of the software and incomplete $D_{it}(E)$ data in the upper half of the bandgap, an effective trap density of $+1.85\times10^{13}\,cm^{-2}eV^{-1}$ was modelled at all energies. The value was derived from the D_{it} data of Figure 7.4 for the condition of threshold. This resulted in a negative shift of V_{th} to $-1.29\,V$ (see Figure 7.5).

In regards to Q_{fixed}, the net charge contribution (magnitude and polarity) is dependent on the thickness of the Al_2O_3 dielectric film [157, 299, 300]. Dielectric films thinner than 5 nm are reported to be comprised of a negative fixed oxide charge, while thicker films present with a positive charge [299]. This thickness-dependent charge character is reconciled on the basis that Q_{fixed} is composed of two charge components of opposite polarities; a negative interfacial oxide charge (Q_{int}) located close to the dielectric/III-V interface and a positive bulk oxide charge (Q_{bulk}) uniformly distributed throughout the dielectric film. This non-uniform distribution of Q_{fixed} arise from the local variation of the film stoichiometry. From XPS investigations Al_2O_3 is found to be oxygen-rich near the interface and aluminium-rich away from it, which in turn correlates to the change in

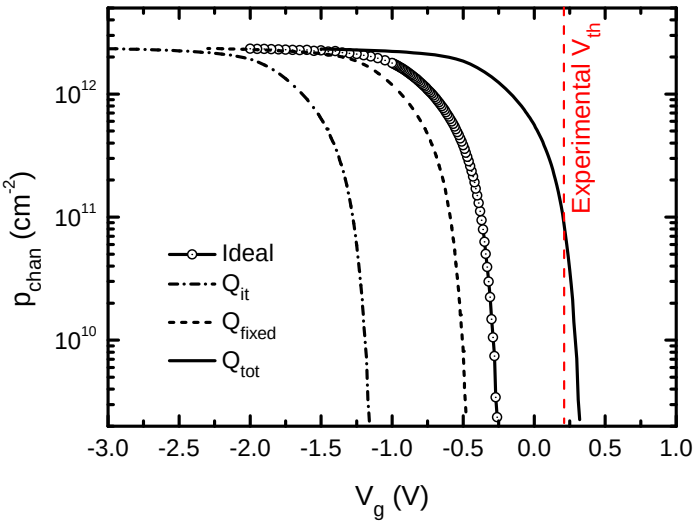


Figure 7.5: Poisson-Schrödinger simulations illustrating the channel carrier density as a function of gate bias for the c802 device wafer under ideal and non-ideal gate stack conditions. Simulations for the following non-ideal charge contributions are shown: interface trap charge (Q_{it}) only, fixed oxide charge (Q_{fixed}) only and total charge (Q_{tot}) comprising Q_{it}, Q_{fixed} and fixed metal/dielectric interface charge ($Q_{m/d}$).

charge polarity across the dielectric film [299]. The results of the simulation taking into account only these fixed charge contributions, experimentally quantified in [157], is shown in Figure 7.5. The resultant threshold voltage ($V_{\mathrm{th}} = -0.59\,\mathrm{eV}$) is shifted negatively from the ideal simulation, similar to that obtained from the interface trap charge analysis. However, from a magnitude perspective, Q_{it} plays the dominant role towards the negative V_{th} shift.

Based on the analysis thus far, the only logical explanation that would account for the experimental V_{th}, that is positively shifted from the ideal case, is the presence of negative fixed charge at the metal/dielectric interface. The $Q_{\mathrm{m/d}}$ charge contribution required to compensate for the positive charges of Q_{it} and Q_{fixed}, and at the same time match the experimental V_{th}, was calculated from 1D-PS simulations. From the matching between the experimental and Q_{tot}-modelled threshold voltage (see Figure 7.5), $Q_{\mathrm{m/d}}$ was evaluated to be $-3.89\times10^{14}\,\mathrm{cm}^{-2}$. Such a charge contribution is likely to arise during gate fabrication from the relevant processes of baking, oxygen plasma ashing and metal deposition. It is possible that the various processes initiate a complex interaction between the dielectric surface, resist residues and gate metal which in turn leaves the metal/dielectric interface populated by a net negative charge.

7.2.2.2 Subthreshold Performance

The ability to minimise power consumption in standby mode is vital to digital logic operation. Integral to this is a low off-state current. An I_{off} of $5.6\,\mu\mathrm{A/mm}$ was determined for the c802 devices based on the earlier mentioned extraction technique for on/off current ratio [295]. Although this value is smaller than the ITRS benchmark of $100\,\mu\mathrm{A/mm}$ [287], the on/off current ratio of 74 suggests leakage issues could be limiting the minimum I_{off} that can be achieved. Drawing on the relevant leakage mechanisms contributing to the off-state current [301], the subthreshold performance of the c802 devices is rationalised in the following paragraphs.

One of the main contributors of off-state leakage in MOSFETs featuring thin gate dielectrics is gate tunnelling (leakage) current. For an assessment of the gate leakage contribution to the off-state performance, the I_{g}–V_{g} characteristic of the c802 device is shown in Figure 7.6. Relating this plot to the subthreshold characteristics (see Figure 7.2), it is seen that in the low drain bias region ($V_{\mathrm{d}} \leq -100\,\mathrm{mV}$) the gate leakage current appears to be limiting the minimum subthreshold current. This however only occurs for $V_{\mathrm{g}} \geq 1.3\,\mathrm{V}$. At the gate bias corresponding to I_{off} ($\sim 0.85\,\mathrm{V}$), the gate leakage current remains a factor of 10 smaller than the subthreshold current. For higher drain biases,

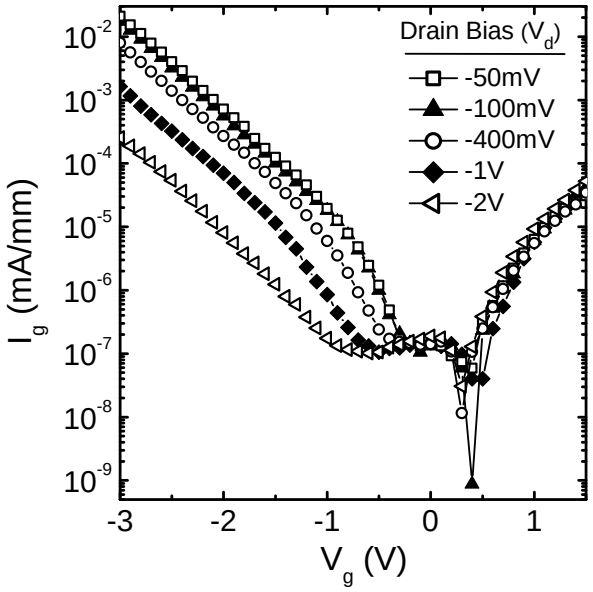


Figure 7.6: Measured I_{g}–V_{g} characteristic of a $2\,\mu$m - L_{g}, $1\,\mu$m - L_{side} p-MOSFET realised on a c802 device wafer.

it is clearly observed that the gate leakage has no impact on the minimum subthreshold current. In fact for $V_{\mathrm{d}} = -2\,\mathrm{V}$ the gate leakage current is more than two orders of magnitude lower than the subthreshold current. From this it appears unlikely the off-state current is affected by the gate leakage mechanism.

Attention is now turned to the channel current that flows between source and drain in the subthreshold region of device operation, referred to as the subthreshold leakage current. From [302] it is noted that the subthreshold transistor performance is a strong function of the subthreshold swing and the drain-induced barrier lowering. Hence, in order to obtain low off-state currents both these device metrics need to be kept to a minimum.

The minimum subthreshold swing of the $2\,\mu$m - L_{g} device was previously highlighted to be $301\,\mathrm{mV/dec}$ in the low drain bias regime. This value is about $3\times$ higher than planar Si MOSFETs which present with swings typically in the range of $70-120\,\mathrm{mV/dec}$ [301]. According to Equation (4.26), three parameters dictate the resulting subthreshold swing in devices. The first is related to the capacitance contributed by the interface trap states (C_{it}). A large C_{it} effectively deteriorates the swing. Therefore, the high swing is likely to arise from the presence of a large number of interface trap states in the region of the bandgap swept by the Fermi level during subthreshold operation. A method that is commonly employed to ascertain the effective interface trap density is the subthreshold slope technique (see Section 4.4.2.4). Applying Equation (4.27), with the SS value obtained from device measurements, results in a D_{it} of $1.47\times10^{13}\,\mathrm{cm^{-2}eV^{-1}}$.

In order to estimate the position of the trap energy corresponding to the SS, the procedure outlined in [173] is used. The first step is to determine the gate voltage at the point of minimum SS. As shown in Figure 7.7(a), this minimum occurs at $V_g = 0.3\,\mathrm{V}$. This bias is then mapped onto the I_d–V_g measurement (see Figure 7.7(b)) enabling the drain current at the point of minimum swing to be evaluated. Also shown in Figure 7.7(b) is a transition point. The intersection of lines extrapolated from the saturation and subthreshold regions define this point. It is noted that the current at the minimum SS decreases by a factor of 43.4 from that at the transition point. This fractional decrease in current is then mapped onto the p_chan–V_g profile shown in Figure 7.7(c). The carrier density in the channel, calculated from 1D-PS simulations [280], serves as a fair approximation of the device current since $I_\mathrm{d} \propto p_\mathrm{chan}$. As before, a transition point is first defined. A change in current by a factor of 43.4 from this transition point gives rise to the point of minimum SS being located at $V_\mathrm{g} = 0.325\,\mathrm{V}$. The trap energy corresponding to this bias, extracted from the E_tr–V_g relationship shown in Figure 7.7(c), was $-0.29\,\mathrm{eV}$. Knowledge of E_tr also enables the trap density at minimum SS to be determined from the $D_{it}(E)$ profile indicated in Figure 7.4 (obtained from [173]). From this, an interface trap density of $1.4{\times}10^{13}\,\mathrm{cm^{-2}eV^{-1}}$ is extracted. It is noteworthy that this value shows good agreement with the D_it extracted at minimum SS from the $2\,\mu\mathrm{m}$-L_g device. This further reiterates the fact that interfaces identically passivated should present with very similar D_it numbers.

The importance of D_it to the subthreshold swing can be appreciated by assuming the trap density at minimum SS is lower than the extracted value by an order of magnitude. This drops the swing by a factor of 3.5 ($SS = 83.7\,\mathrm{mV/dec}$). There is evidence to show that D_it approaches values of $1{\times}10^{12}\,\mathrm{cm^{-2}eV^{-1}}$, or lower in the upper half of the bandgap towards the conduction band for sulphide passivated Al_2O_3/$In_{0.53}Ga_{0.47}As$ interfaces [297, 298]. While D_it numbers will vary depending on the adopted passivation technique, it is reasonable to infer that trap densities will be lower closer to the conduction band for an Al_2O_3/$In_{0.53}Ga_{0.47}As$ interface subject to arsenic capping/de-capping based on the descending D_it profile shown in Figure 7.4. Hence, shifting the window of energy through which the transistor operates towards the conduction band edge, should enable much improved SS to be realised. This can be attained by using a gate metal of a higher work function. In the process, the on-state performance would also see improvement since the Fermi level will sweep through lower D_it values, thereby enhancing gate control of the channel.

Another parameter that impacts SS is the gate dielectric capacitance (C_ox). From

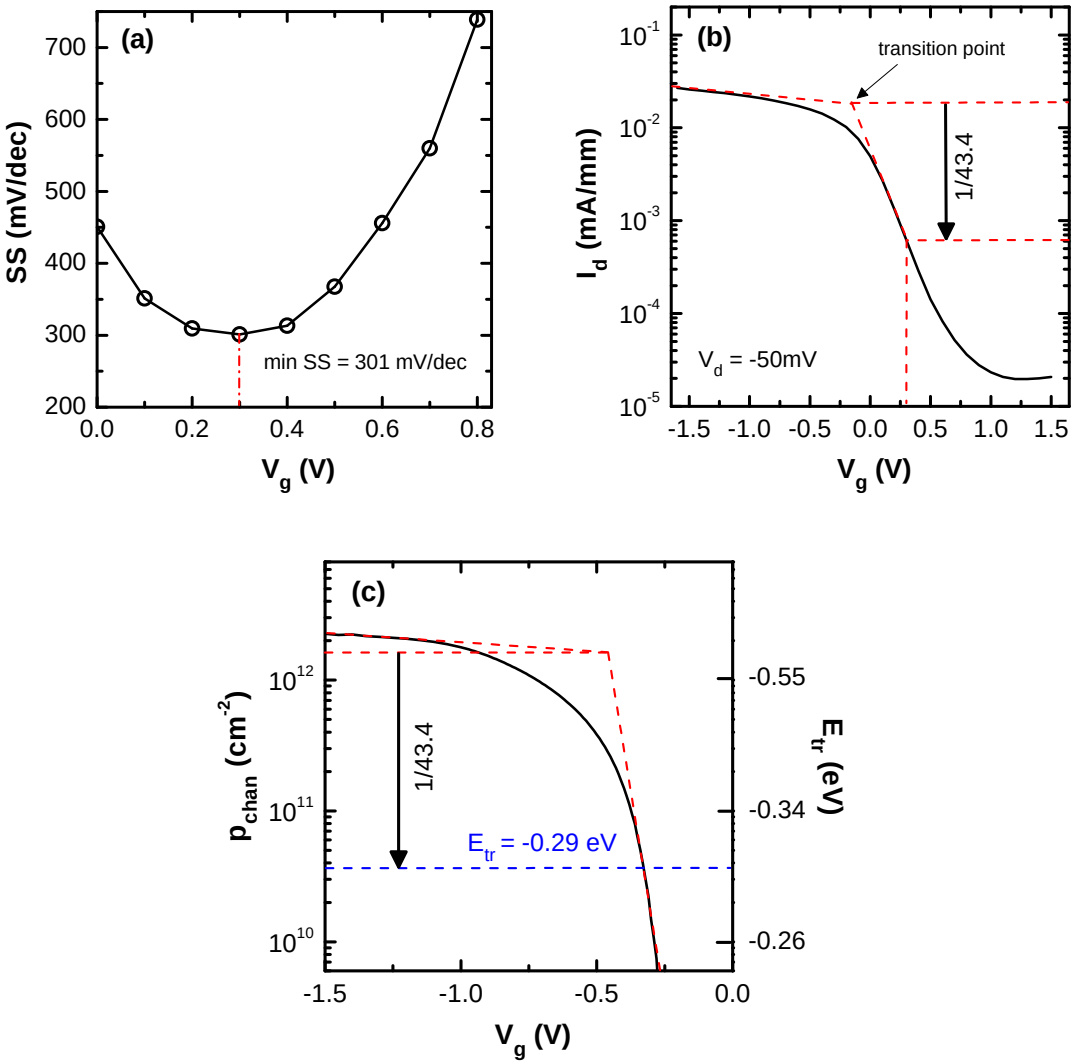


Figure 7.7: Gate bias dependence of (a) the subthreshold swing, (b) the drain current at $V_\mathrm{d} = -50\,\mathrm{mV}$ and (c) the channel carrier density and trap energy calculated from Poisson-Schrödinger simulations for the c802 device wafer. Trap energy (E_tr) is defined as zero at the conduction band edge and negative within the bandgap.

Equation (4.26), it is observed that increasing this capacitance would improve the swing. The reason for this is a higher dielectric capacitance translates to a lower EOT and hence, a lower CET, which in turn improves the gate control of the channel. This can be realised by reducing the thickness of the dielectric and/or employing a dielectric with a higher permittivity. By way of the former, a SS of 211.7 mV/dec (approximately 30% improvement) can be achieved by scaling the physical thickness of the Al_2O_3 film from 6.5 nm (EOT = 3.17 nm) to 2 nm (EOT = 0.98 nm) for instance. However the trade-off with thickness scaling is an increase in gate leakage which ultimately sets the minimum

thickness the dielectric can be scaled to. Alternatively, switching to higher-k dielectrics would allow for EOT scaling without compromising on leakage performance. An EOT of 0.98 nm, similar to the 2 nm Al_2O_3 film, can be obtained from HfO_2 ($k = 20$) with a physical film thickness of 5 nm, which, in theory, should result in a similar swing. The composition of the gate dielectric, though, has a direct bearing on the quality of the high-k/semiconductor interface. HfO_2/InGaAs systems have been reported to have an inferior interface quality, with D_{it} increasing by as much as two orders of magnitude in relation to Al_2O_3/InGaAs [23, 35, 38, 303, 304]. This however can be mitigated by employing a bilayer dielectric, whereby Al_2O_3 is employed as an interface control layer between the semiconductor and a higher-k dielectric. Good quality interfaces have been demonstrated in high-k/Al_2O_3/InGaAs systems featuring HfO_2 [23, 303–305], TiO_2 [30] and ZrO_2 [184]. The adoption of bilayer dielectrics would therefore enable further EOT scaling for improvement of SS, without compromising on either the interface quality or the leakage floor. An important point to note here is that D_{it} has an overriding bearing on the outcome of the subthreshold swing in comparison to the dielectric capacitance.

Apart from C_{ox} and C_{it}, the subthreshold swing is also influenced by the depletion layer capacitance (C_d). A lower swing is obtained by minimising C_d (see Equation (4.26)). To a first approximation, the depletion layer capacitance is proportional to the capacitance of the quantum well channel (ε_s/t_{chan}). Therefore, increasing the channel thickness would decrease C_d, thereby improving the swing. However this comes at the cost of higher DIBL, which is the other metric affecting the subthreshold drain current. DIBL of the 2 μm L_g device is calculated to be 46.7 mV/V, based on the threshold voltages determined at drain biases of -50 mV and -2 V. This small value is suggestive of a well-scaled device layer structure at this gate length for which SCE is negligible.

7.2.2.3 Parasitic S/D Series Resistance

Shown in Figure 7.8 is the on-resistance, extracted at a gate bias of -3 V and a drain bias of -50 mV, as a function of transistor gate length. From the R_{on} versus L_g plot the total source/drain series resistance, determined from the intercept of the fitted line with the y-axis, is 732 Ω.mm. This high R_{sd} value could arise from the ohmic contact resistance and/or the access resistance between the gate edge and the S/D contacts. In order to elucidate the specific contributions of these resistance components to the overall S/D resistance, both R_c and R_{side} were experimentally quantified. From TLM measurements, the contact resistance was evaluated to be 6.58 Ω.mm. The access resistance, R_{side}, calculated from Equation (6.4) using the aforementioned values of R_{sd} and R_c, is found

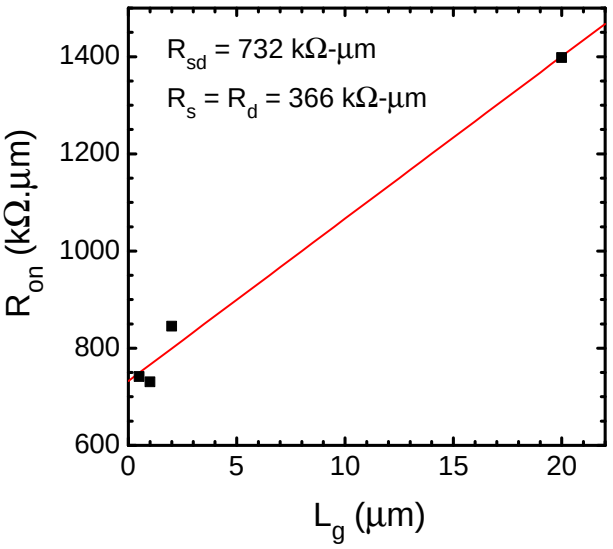


Figure 7.8: R_{on} as a function of gate length for c802 devices, measured at $V_g = -3\,V$ and $V_d = -50\,mV$.

to be 359.42 Ω.mm. It is seen that R_{sd} is dominated by the access resistance with only a 2% contribution from the contact resistance.

Two aspects can be drawn on to account for the dominance of access resistance to R_{sd}. The first is the lateral spacing from the edges of the gate to the S/D contacts. Transistors built on c802 feature a L_{side} of 1 μm. Therefore, by downscaling L_{side} the access resistance should be drastically reduced. For instance, a ten-fold reduction in R_{side} can be expected by scaling L_{side} gap to 100 nm. However the resulting access resistance (35.9 Ω.mm) would still account for 84.5% of the total series resistance. The large access resistance contribution to R_{sd}, even at such a scaled L_{side} gap, is attributed to the high sheet resistance of the grown material. From the R_{side} relationship given in Equation (6.4), R_{sh} was evaluated to be 359.42 kΩ/sq. Reasons for the high sheet resistance of the c802 device wafer will be discussed in the next section. Apart from the extrinsic device performance being negatively impacted by the parasitic losses arising from R_{side}, the high sheet resistance would also affect the intrinsic device transport by limiting the maximum achievable drive current. Moreover, it is to be expected that the large parasitic series resistance would significantly affect the on-current in shorter gate length devices.

7.2.2.4 Drain Current and Transconductance

It was highlighted in Section 7.2.2.3 that devices fabricated on the c802 wafer suffer from a high source/drain series resistance, which in turn limits the on-current. However this alone

does not account for the poor on-state device characteristics. In the following paragraphs the other performance-limiting factors are examined.

The transfer characteristics of the device in the linear and saturation regions are shown in Figure 7.9. On inspecting the linear region characteristics (see Figure 7.9(a)), it is notable that the drain current does not show a complete linear response in the investigated gate bias range. In the low gate bias regime, the drain current increases approximately linearly as expected from the I_{d}–V_{g} relationship for $V_{\mathrm{d}} < V_{\mathrm{g}} - V_{\mathrm{th}}$ (see Equation (2.5)). However at higher gate biases the drain current begins to show a sub-linear increase with V_{g}. Similar observations of drain current dependence on gate bias have been previously reported [182, 306]. This trend is ascribed to two factors [95], the first relates to the dependence of the effective mobility on the transverse electric field across the gate region. With increasing gate bias the mobility degrades, thereby causing the observed sub-linear drain current response. An analysis of the resulting mobility will be dealt with in Section 7.2.2.5. Source/drain series resistance is the other cause for the sub-linear I_{d}–V_{g} characteristic. Depending upon the gate bias, part of the applied drain bias (extrinsic) is lost to device parasitics. As a result of the voltage loss associated with R_{sd}, the drain bias seen by the channel in the gated region (intrinsic) is reduced from the applied bias. This in turn causes a sub-linear increase of I_{d} with respect to V_{g}. Interestingly at even higher gate biases the sub-linear response disappears and the I_{d}–V_{g} characteristic reverts back to being linear again. Similar to the I_{d}–V_{g} relationship

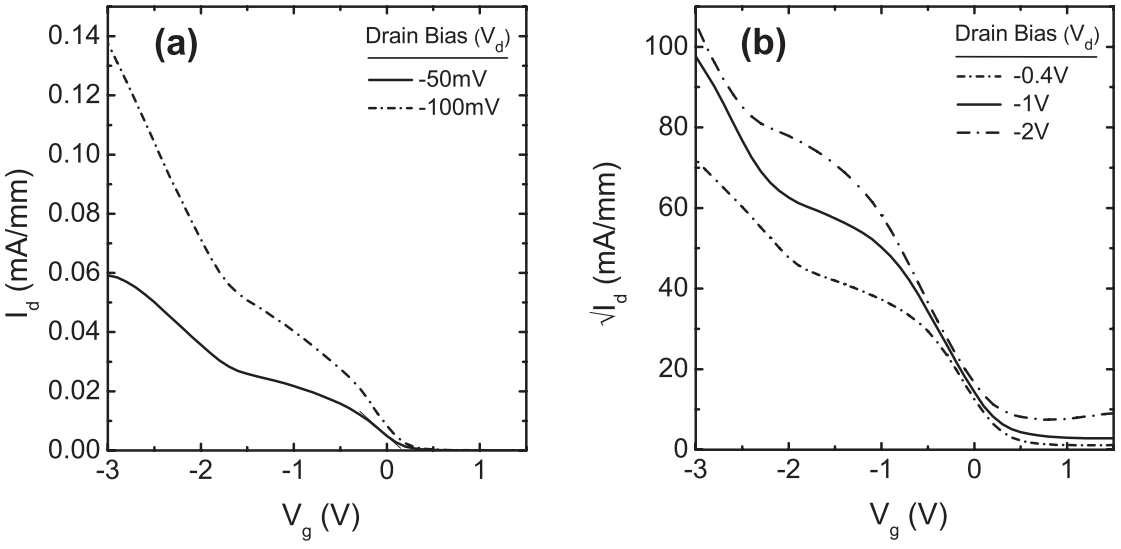


Figure 7.9: Transfer characteristics of the c802 device with $L_{\mathrm{g}} = 2\,\mu$m and $L_{\mathrm{side}} = 1\,\mu$m in (a) the linear region plotted as I_{d} versus V_{g} and (b) in the saturation region plotted as $\sqrt{I_{\mathrm{d}}}$ versus V_{g}.

in the linear region of device operation, a linear drain current response is also expected from the $\sqrt{I_d}$–V_g relationship (see Equation (2.6)) in saturation ($V_d > V_g - V_{th}$). A plot of the square root of the drain current versus gate bias in saturation is illustrated in Figure 7.9(b). From this it is observed that the transfer characteristic in the saturation region is consistent with the behaviour reported for the c802 devices in the linear region of operation.

The appearance of two linear responses, interposed by a sub-linear increase of the drain current, in the transfer characteristics is also manifested as two g_m peaks (see Figure 7.2(c)). This occurrence is associated with the formation of two conducting channels in the device [307–309]. Illustrated in Figure 7.10 is the band diagram of the c802 wafer. It is noted that apart from the buried $In_{0.75}Ga_{0.25}As$ QW, a second quantum well is also present at the Al_2O_3 gate dielectric/InGaAs cap interface. This surface QW is a direct result of the band lineup between the cap and barrier layer, for which a valence band offset exists between $In_{0.53}Ga_{0.47}As$ and $In_{0.52}Al_{0.48}As$. The implication of the surface quantum well to the resulting I_d–V_g and g_m–V_g characteristics can be appreciated by considering the carrier population during device operation. Since this will be covered during the mobility discussion (see Section 7.2.2.5), the main conclusions are presented here without going into detail. When the device initially turns on, carrier population is confined to the buried $In_{0.75}Ga_{0.25}As$ QW channel and is maintained up to a certain gate bias. The first transconductance peak is therefore related to current conduction in the buried channel. However, at higher gate bias, the band bending also begins to energetically favour

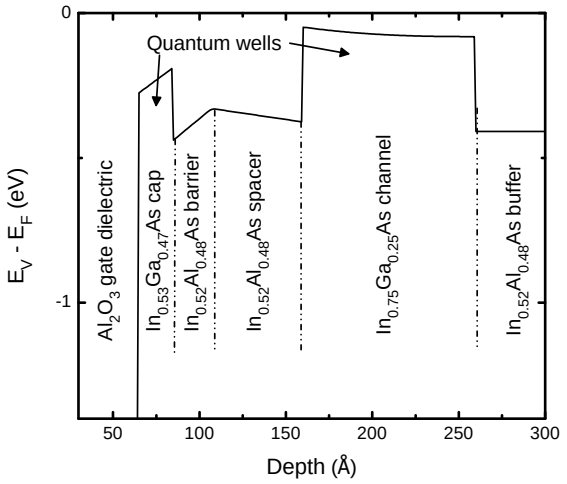


Figure 7.10: Energy band diagram of the c802 device wafer depicting the presence of two quantum wells; a buried $In_{0.75}Ga_{0.25}As$ QW and a surface $In_{0.53}Ga_{0.47}As$ QW.

carrier transfer from the δ-doped supply layer into the surface QW. As the carrier population in the surface channel builds up, the 2DHG in the buried channel begins to be electrostatically shielded from the gate electrode. As a result the carrier density in the buried channel saturates. Further increase in gate bias will only benefit carrier confinement in the surface quantum well. The mobility of these surface holes will be lower due to dielectric and interface-related scattering phenomenon. However, transconductance is determined by the product of the C_{gc} and the μ_{eff}. The transition from buried channel to surface channel device operation gives rise to a larger gate-to-channel capacitance. It is likely the reduction in effective mobility, with increasing gate bias, is superseded by the increased C_{gc} [307, 308]. This in turn gives form to the second g_{m} peak observed at the higher gate bias for which surface channel device operation ensues. The parasitic surface channel and its associated effects can be alleviated through appropriate band engineering. This involves scaling down the thickness of the cap layer to a few monolayers. Enhanced quantum confinement in the surface quantum well, arising from the down scaling, would cause the lowest hole and electron subband energies to increase in energy and move away from the band edges, thereby circumventing carrier spillover [310].

Apart from the double g_{m} peaks, the appearance of kinks in the C–V characteristics is another indicator of carrier spillover into the surface quantum well during device operation. Illustrated in Figure 7.11 are the multi-frequency C–V curves obtained from measurements of annular C–V capacitors fabricated alongside transistors on c802. In agreement with previous experimental observations [307, 311–313], kinks are found to be present in the C–V curves. These kinks are explained by the abrupt increase in the gate-to-channel capacitance brought on by the onset of carrier occupation in the surface quantum well. In theory C_{gc} should approach that of the gate dielectric capacitance, which for Al_2O_3 is $1.1\,\mu\mathrm{F/cm^2}$ based on the physical dielectric thickness of 6.5 nm based on TEM imaging and a dielectric constant of 8 [26]. It is seen that C_{max} at $V_{\mathrm{g}} = -2.5\,\mathrm{V}$ approaches this value for the measurement frequency of 1 kHz. The effects of DOS and charge quantisation [26] can be drawn on to account for the slight deviation of C_{max} from C_{ox}, which is quantified to be $0.05\,\mu\mathrm{F/cm^2}$.

It is further observed that the C–V characteristics of Figure 7.11 present with large frequency dispersion in the range $-2.5\,\mathrm{V} < V_{\mathrm{g}} < 0.5\,\mathrm{V}$. One possible cause of such dispersion behaviour is often related to series resistance [16]. Shown in Figure 7.12(a) is the equivalent circuit of a MOS capacitor consisting of the effective capacitance (C), which is the oxide capacitance (C_{ox}) in series with the semiconductor capacitance (C_{s}), and a series resistance (R). For simplicity the MOS capacitor is assumed to be free

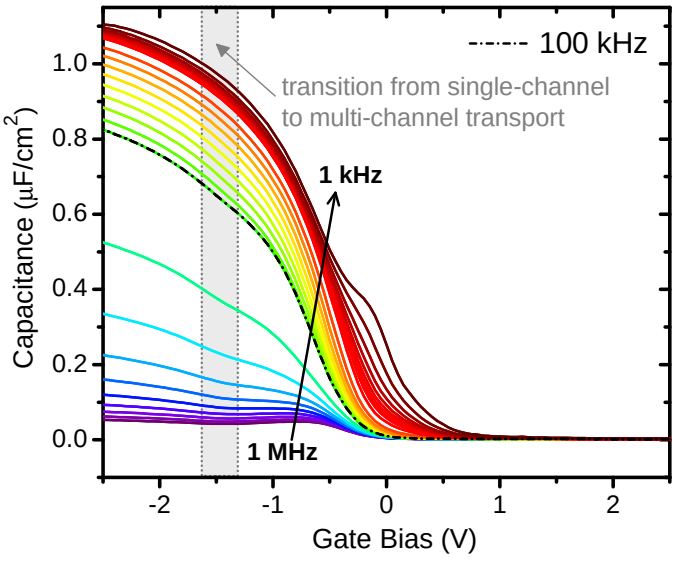


Figure 7.11: C–V (1 kHz to 1 MHz, 295 K) characteristics representative of the c802 structure. MOS capacitors were fabricated using the annular C–V layout featuring 100 μm diameter gate contacts. To provide an accurate measure of device admittance characteristics MOS capacitors were realised alongside MOSFETs on the same sample, and therefore subjected to identical processing conditions.

of interface traps and gate leakage. Also shown in Figure 7.12(a) is the measurement circuit model, which comprises the measured equivalent parallel capacitance (C_{m}) and conductance (G_{m}). The measured capacitance is given by [156, 314]:

$$C_{\mathrm{m}} = \frac{C}{1 + (\omega C R)^2}\,, \tag{7.2}$$

where $\omega = 2\pi f$ (f is the measurement frequency). From this equation it is seen that the measured capacitance is reduced by the series resistance, resulting in a vertical-type shift of the C–V curves with frequency, with its effect on capacitance being strongest in accumulation and at high frequencies. This also illustrated in Figure 7.12(b), in which a simulated ideal, multi-frequency C–V characteristic including the effect of a 3 kΩ series resistance [314] is presented. Series resistance can arise from a number of sources including substrate resistance, contact resistance and access resistance. In the annular C–V layout the gap separation between the circular gate and concentric annular ring (ohmic contact) enclosing it is 10 μm. This separation, coupled with the large sheet resistance of the c802 wafer, gives rise to an access resistance of 3.59 kΩ.mm. Taking the contribution of R_{c} into account, the resulting S/D series resistance is calculated to be 3.6 kΩ.mm. Although this appears to suggest the capacitance measurements are affected by series resistance, there are a number of reasons why series resistance cannot be held accountable for the observed

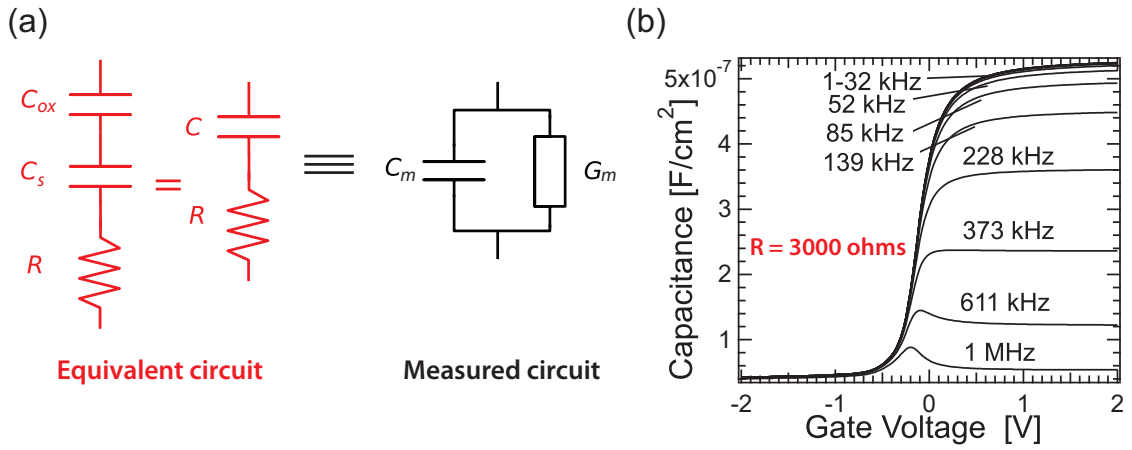


Figure 7.12: (a) Equivalent circuit of a MOS capacitor with series resistance and the measured circuit of a MOS capacitor, and (b) model C–V (1 kHz to 1 MHz) characteristics of a p-MOS capacitor with only series resistance [314].

dispersion.

Frequency dispersion due to series resistance depends on ω^{-2} (see Equation (7.2)). This however is not observed in the experimental results shown in Figure 7.11. In addition to the vertical dispersion, a second type of dispersion, whereby the C–V curves shift horizontally with gate bias, is also observed for $-2.5\,\mathrm{V} < V_\mathrm{g} < 0.5\,\mathrm{V}$. This horizontal-type dispersion (frequency-dependent flatband voltage shift) is not due to series resistance since its effect on the C–V characteristics is limited to a frequency-dependent capacitance change alone (see Figure 7.12(b)). Instead, the observed dispersion is related to interface trap behaviour [163, 314]. Martens modelled the effects of interface traps on the C–V response of a p-MOS capacitor in the absence of series resistance, the results of which are illustrated in Figure 7.13 [314]. It is observed that the C–V response does not exhibit a frequency-dependent flatband shift for $D_\mathrm{it} = 3.8 \times 10^{11}\,\mathrm{cm}^{-2}\mathrm{eV}^{-1}$. However when D_it increases by an order of magnitude, the frequency-dependent flatband shift appears in the C–V characteristics. This occurs when $D_\mathrm{it} > C_\mathrm{ox}/q$ [163, 314] which is the case for $D_\mathrm{it} = 3 \times 10^{12}\,\mathrm{cm}^{-2}\mathrm{eV}^{-1}$ based on the oxide parameters used in the simulation. Martens attributes such frequency-dependent flatband voltage shift to weak Fermi level pinning (WFLP) for which the Fermi level movement although not completely obstructed, is heavily burdened by the presence of interface traps [163, 314]. It is further observed that at even higher trap densities of $D_\mathrm{it} = 3 \times 10^{14}\,\mathrm{cm}^{-2}\mathrm{eV}^{-1}$ a vertical-type frequency dispersion emerges resulting in a frequency dependent capacitance change. Extreme stretch-out can be drawn on to explain this phenomenon [163, 314]. With increasing D_it the stretch-out increases, resulting in very little change of the interface trap time constant, and therefore

Fermi level, with voltage at the highest trap density. This is indicative of strong Fermi level pinning, which explains the additional decrease in capacitance with frequency at the highest trap densities.

It is observed that the measured result shown in Figure 7.11 resembles the simulation result for $D_{\mathrm{it}} = 3 \times 10^{14}\,\mathrm{cm}^{-2}\mathrm{eV}^{-1}$ shown in Figure 7.13. In the case of c802 structure, WFLP starts to occur when $D_{\mathrm{it}} > 6.9 \times 10^{12}\,\mathrm{cm}^{-2}\mathrm{eV}^{-1}$. Referring back to the D_{it} distribution of the $Al_2O_3/In_{0.53}Ga_{0.47}As$ illustrated in Figure 7.4, it is seen that the trap density increases by almost an order of magnitude to $1{\times}10^{14}\,\mathrm{cm}^{-2}\mathrm{eV}^{-1}$ over a limited energy range of 0.15 eV after device turn on. Such high trap densities correlate well with the horizontal and vertical frequency dispersion noted in the C–V curves. Moreover, the small C–V hump observed at $V_{\mathrm{g}} = 0\,\mathrm{V}$ is a typical signature of interface trap behaviour [315]. Due to the high density of interface traps in the region of the bandgap swept by the Fermi level above threshold, the efficiency of the gate control is reduced. This partly

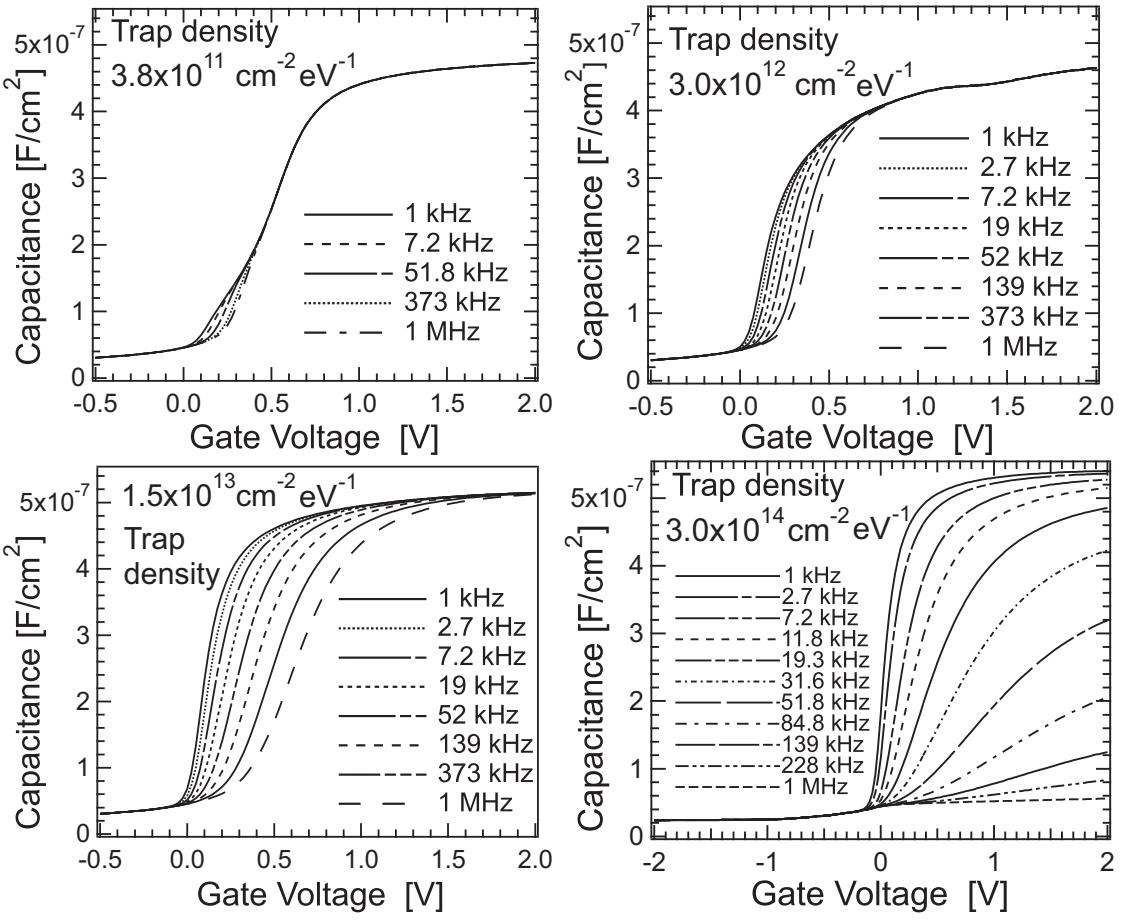


Figure 7.13: Simulations illustrating the effect of increasing interface trap density on the C–V (1 kHz to 1 MHz, 295 K) characteristics of a p-MOS capacitor [314].

accounts for the poor on-state performance.

The gate modulation efficiency is further hampered by the large gate-to-channel distance, which acts to reduce C_{gc}. As a consequence the transconductance and the resulting drain current is lowered. Scaling down the cap layer for elimination of parasitic surface conduction would improve the gate-to-channel capacitance. In addition, C_{gc} would also benefit from thinning down the gate dielectric thickness. One limitation with the adoption of Al_2O_3 is its dielectric constant ($\epsilon_{ox} = 8$). Switching to higher-k dielectrics such as HfO_2 would also help increase the gate-to-channel capacitance. A final improvement in on-state performance can be achieved by adopting the inverted-channel design concept, which eliminates the additional spacer layer thickness above the channel. The c805 devices, discussed in Section 7.2.3, are based on such a design.

7.2.2.5 Mobility Analysis

Effective mobility is a key parameter in assessing device performance. Illustrated in Figure 7.14 is the effective hole mobility of the c802 device extracted as a function of carrier density using a combination of C–V and I–V measurements (see Section 4.4.2.3). The parameter extraction was based on the 100 kHz measurement (see Figure 7.11) of a 100 μm-diameter capacitor and the linear I_d–V_g characteristics ($V_d = 50$ mV) of a 20 μm-L_g MOSFET.

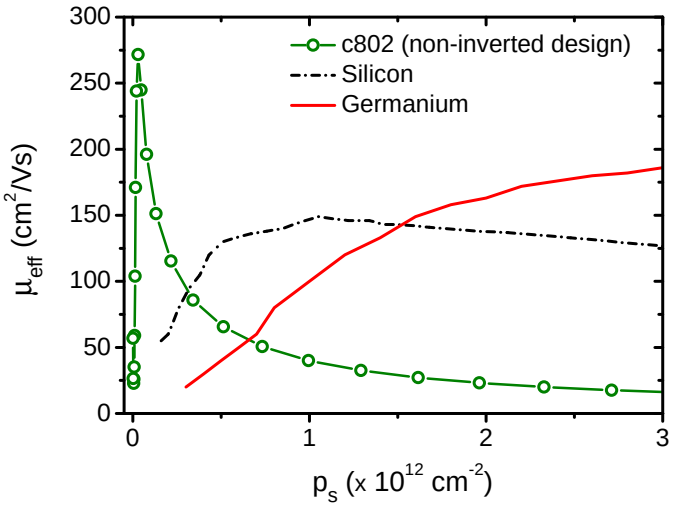


Figure 7.14: Effective hole mobility versus sheet charge density for the c802 device, extracted using a combination of the I_d–V_g measurement of a 20 μm-L_g transistor at $V_d = -50$ mV and C–V measurement of a 100 μm-diameter MOS capacitor at 100 kHz. Included for comparison is the universal hole mobility of Si [316] and Ge [92].

The extracted μ_{eff}–p_{s} data of the c802 device presents with a peak effective mobility of 271 cm^2/Vs. Universal hole mobility in Ge and Si is also plotted for comparison (see Figure 7.14). It is observed that the peak mobility of the c802 MOSFET is 1.85× and 1.4× higher in comparison to Si and Ge respectively. However, this peak occurs at a relatively low carrier density of $2.96\times10^{10}\,\text{cm}^{-2}$, and the mobility gain over Si and Ge is not maintained at higher densities. Beyond the aforementioned carrier density value, the mobility of the c802 device decreases rapidly reaching a value of 40 cm^2/Vs at $p_{\text{s}} = 1\times10^{12}\,\text{cm}^{-2}$. It is notable that at this carrier density μ_{eff} of c802 is 72% lower compared to Si and 60% lower than in Ge. With increasing carrier density, c802 presents with a more gradual reduction in mobility. Note that the mobility was extracted, with corrections for parasitic S/D series resistance applied, but without interface trap corrections. Inclusion of the latter correction during the extraction should improve the resulting μ_{eff} values. The reason for this is that the capacitive contribution of the interface traps to the measured capacitance leads to an overestimation of charge, which in turn results in the effective mobility being underestimated [186, 317–319]. Alternatively, mobility calculation based on the Hall effect technique [318] or capacitance measurement at low temperatures [26, 319] would alleviate the D_{it} response, enabling a more accurate μ_{eff} to be evaluated.

To shed light on the μ_{eff}–p_{s} relationship of the c802 device, 1D-PS simulations were undertaken to map the hole population at different regions of device operation and are shown in Figure 7.15. Below threshold the entire channel is populated by holes (see Figure 7.15(a)). As a result, the mobility is limited by the effects of interface scattering from both the back and the front interfaces of the buried channel, as well as remote ionised impurity scattering from the δ-doping. For $V_{\text{g}} > V_{\text{th}}$, the charge centroid is located between the centre and the front interface of the channel. Nonetheless, a significant portion of the holes still populate the back of the channel (see Figure 7.15(b)). As holes remain populated in the entire channel, the previous mentioned scattering mechanisms are still in play. However the increase in carrier density enhances the screening of the dopants. This in turns causes the mobility to rise and peak above threshold. With increasing gate bias the charge centroid moves closer to the front of the channel, with an increase in the hole concentration (see Figure 7.15(c)). In this region of operation the mobility becomes dominated by interface scattering from the front interface of the channel, while the effects of impurity scattering are still effectively screened. As a result the mobility starts to show a decline. In addition, a small number of holes start to populate the quantum well above the delta-doping resulting in a parasitic surface channel to form (see Figure 7.15(c)). Further increase in gate bias causes the hole concentration to rise with the charge centroid located approximately at the centre of the parasitic channel (see Figure 7.15(d)). Since

no strain is induced in this quantum well, it is to be expected that the holes in this parasitic channel would have a significantly lower mobility in comparison to the buried channel. Additionally, due to multi-channel transport, μ_{eff} becomes a weighted average of the carrier hole mobility in the buried and surface channels [312].

From a scattering perspective, the thinner channel and the closer proximity of the δ-doping layer translate to interface scattering (due to front/back channel interface) and ionised impurity scattering having a greater impact on surface channel mobility. The proximity of the channel to the high-k dielectric also means surface holes are more strongly affected by remote Coulomb scattering (RCS) from fixed charges within the dielectric and/or trapped charges at the dielectric/semiconductor interface [81, 312]. There is also the additional scattering mechanism from interface roughness [81]. These effects combined

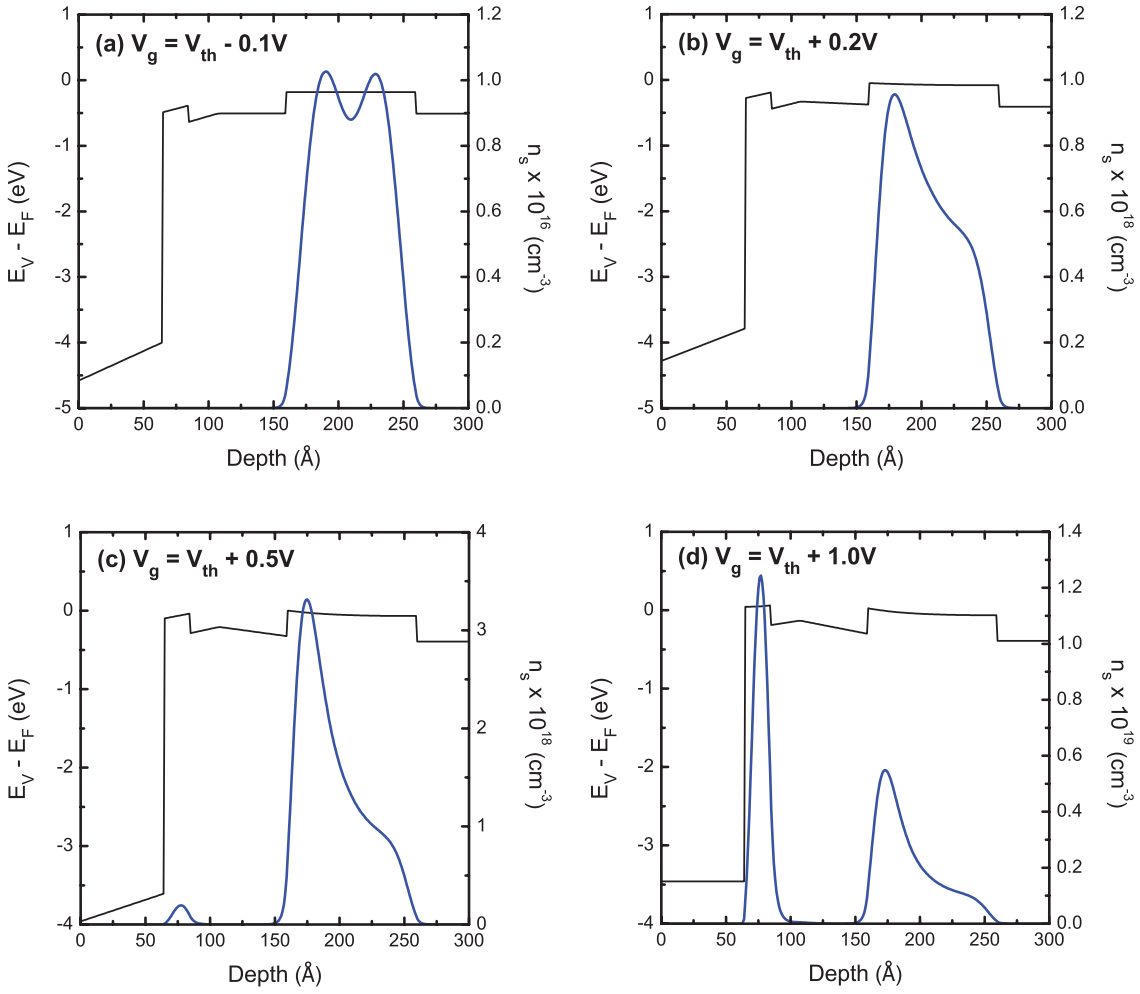


Figure 7.15: One-dimensional Poisson-Schrödinger simulations of hole populations of the c802 device for differing gate bias conditions.

give rise to the observed degradation of hole mobility at higher carrier densities.

The ability to achieve and maintain a high effective mobility at higher densities is correlated to the elimination of the parasitic surface conduction path. As previously mentioned, this can be attained by thinning down the cap layer to a few monolayers in thickness [310], which in turn would ensure only buried channel operation is sustained at all gate biases. Moreover, increasing the strain by moving to higher indium-content channels would enable much higher effective mobilities to be realised in the range of device operation.

7.2.3 c805 Device Results and Discussion

In this section the performance of a $2\,\mu$m-L_g device at a fixed gate-to-source/drain separation is first discussed. Following this, an assessment of the impact of L_{side} gap scaling on device performance is presented.

7.2.3.1 $2\,\mu$m-L_g Device Performance

Figure 7.16 illustrates the DC output and transfer characteristics of a device with a L_g of $2\,\mu$m and a L_{side} of $1\,\mu$m, realised on a c805 wafer. These characteristics are representative of measured devices of identical geometries.

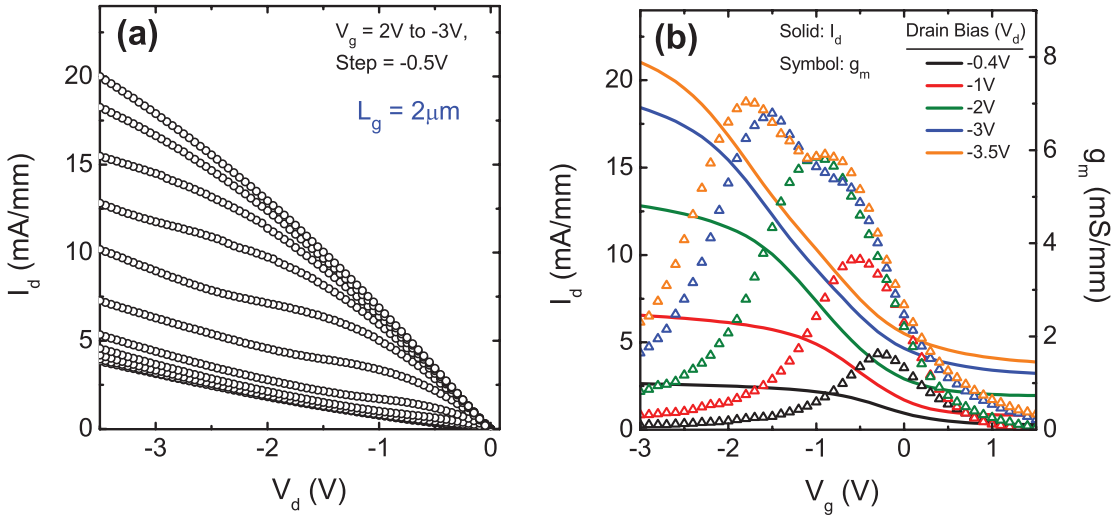


Figure 7.16: Measured (a) output and (b) transfer characteristics of $2\,\mu$m-L_g, $1\,\mu$m-L_{side} p-MOSFET realised on a c805 device wafer.

From the output characteristics it is observed that the c805 device demonstrates control of the drain current by the gate voltage. However the drain current continues to increase at higher drain biases, without showing saturation. A maximum $I_{\mathrm{d,sat}}$ of 12.8 mA/mm is obtained at $V_{\mathrm{g}} = -3\,\mathrm{V}$ and $V_{\mathrm{d}} = -2\,\mathrm{V}$. Based on the transfer characteristics a peak g_{m} of 5.85 mS/mm is extracted at $V_{\mathrm{d}} = -2\,\mathrm{V}$. It is noteworthy that the device fabricated on the c805 wafer presents with a significant leap in on-state performance compared to previous results on the c802 device (see Figure 7.2), subject to similar processing and featuring identical geometrical layouts. For the same biases conditions c805 shows a 11× enhancement in drive current and 5.5× improvement in transconductance over that of the c802 device. From the I_{d}–V_{d} plot, a R_{on} of 151 kΩ.μm was extracted at $V_{\mathrm{g}} = -3\,\mathrm{V}$ and $V_{\mathrm{d}} = -50\,\mathrm{mV}$. In comparison to c802, the on-resistance of c805 is 5.6× lower, however there is a clear trade-off between the on-state and off-state performance. It is noticeable that the c805 device is plagued by significant subthreshold leakage. As a result, unlike the good switch-off characteristics of the c802 devices, devices on c805 do not switch-off. Furthermore, the off-state performance degrades with increasing drain bias due to higher subthreshold leakage currents. At $V_{\mathrm{g}} = 1.5\,\mathrm{V}$, the leakage is found to increase by a factor of 50 from 38.1 μA/mm at $V_{\mathrm{d}} = -50\,\mathrm{mV}$ to 1.95 mA/mm at $V_{\mathrm{d}} = -2\,\mathrm{V}$. The significance of the subthreshold leakage is underscored by the fact that the leakage current is approximately 15% of the maximum drain current at $V_{\mathrm{d}} = -2\,\mathrm{V}$.

Off-state Performance

The aim of this discussion is to examine the possible causes that would explain the poor off-state device characteristics. The interrogation begins with the consideration of the gate leakage effects on the subthreshold drain current. For this purpose, a semi-logarithmic plot of the I_{d}–V_{g} and I_{g}–V_{g} curves are shown in Figure 7.17. It is observed that the gate leakage current is at least three orders of magnitude smaller than the drain current in the gate bias range of 0 V to 1.5 V. From this it is concluded that gate leakage is not the reason for the high subthreshold currents.

A common source of poor off-state performance, resulting in incomplete device switch-off, is a parasitic leakage path below the channel. This can arise from the delta-doping layer behaving as a parasitic hole conduction layer. However, 1D-PS simulations did not reveal any apparent hole population in the delta-doped layer (see Figure 7.19). Another source of parasitic conduction often encountered in devices is attributed to leakage through the buffer layer [91, 320, 321]. Buffer conduction mainly originates from non-intentionally introduced p-type impurities and/or defects that are acceptor-like in nature during MBE growth, which act as hole carriers [321]. This parasitic conduction can be suppressed by

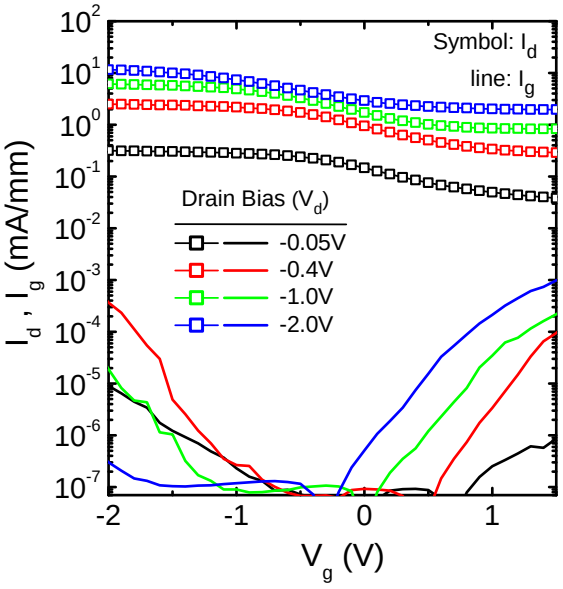


Figure 7.17: Semi-logarithmic plot of the measured I_{d}–V_{g} and I_{g}–V_{g} characteristics of the 2 μm - L_{g}, 1 μm - L_{side} p-MOSFET realised on a c805 device wafer.

employing n-type doping in the buffer, either as bulk or delta doping [69, 321], thereby effectively isolating the channel from the buffer layer. In evidence to this, Irisawa *et al.* [321] demonstrated a factor of 40 reduction in the off-state current of Ge p-channel MOSFET by introducing antimony (n-type) delta doping 100 nm below the channel. Additionally, the effects of BTBT in narrow gap semiconductors cannot be ignored. Given the bandgap of the $In_{0.75}Ga_{0.25}As$ channel is only 0.54 eV, the leakage due to BTBT is partly held accountable for the observed high subthreshold currents.

One phenomenon that would explain the poor switch-off characteristics is Fermi level pinning. This arises from the presence of a large density of interface traps at energy positions coincident with the Fermi level sweep during device operation. Due to significant carrier trapping, the movement of the Fermi level becomes blocked at specific energy positions. As a consequence the carrier density in the channel can no longer be modulated by the gate potential. Depending on the pinning position, either the off-state or on-state device performance is affected. If pinning occurs as the device attempts to turn off then the channel would remain populated with carriers, resulting in incomplete switch-off. Based on the D_{it} profile illustrated in Figure 7.4, FLP is highly likely to occur in the lower half of the bandgap closer to the valence band edge, where trap densities reach values of $1{\times}10^{14}\,\mathrm{cm}^{-2}\mathrm{eV}^{-1}$ and show an ascending trend towards E_v. There is also experimental evidence demonstrating FLP within the conduction band. Using a combination of Hall

measurements and the split C–V technique, Taoka *et al.* [322, 323] quantitatively determined the D_{it} distribution within the conduction band of high-k/InGaAs MOS system. The D_{it} profile was found to be dependent on the choice of gate dielectric, the interface passivation technology and indium composition of InGaAs. However the D_{it} distribution appeared to be relatively insensitive to the thickness of the gate dielectric. The pinning position was evaluated to occur at trap energies in the range of 0.21 eV to 0.35 eV above the conduction band edge of the $Al_2O_3/In_{0.53}Ga_{0.47}As$ system [322, 323]. Based on 1D 1D-PS calculation, trap energies associated with c805 device operation is plotted in Figure 7.18. It is seen that in the subthreshold region the Fermi level sweeps trap energies greater than 0.2 eV. This falls within the aforementioned trap energy range associated with FLP. Therefore, it is highly likely FLP is preventing channel modulation in c805 as the device is biased towards turn-off.

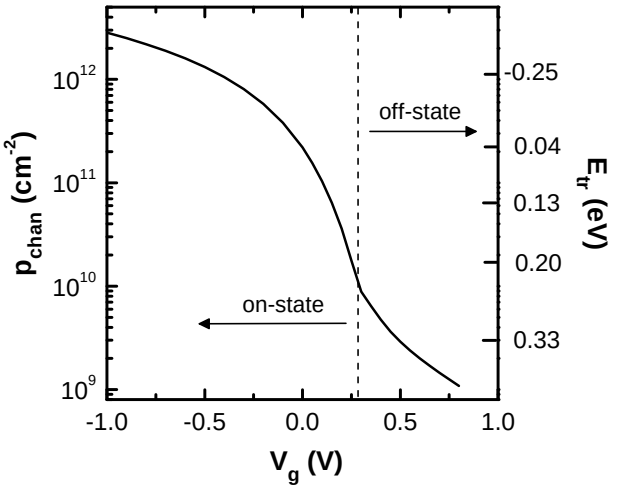


Figure 7.18: Plot of channel carrier density and trap energy as a function of gate bias for the c805 device wafer based on Poisson-Schrödinger simulations. Trap energy (E_{tr}) is defined as zero at the conduction band edge and negative within the bandgap.

An additional mechanism that could be hampering device turn-off is related to the screening of gate potential due to electron accumulation in the $In_{0.53}Ga_{0.47}As$ cap layer [113]. In subthreshold operation the Fermi level moves into the conduction band (see Figure 7.18). This in turn results in the parasitic surface quantum well being populated by electrons. With increasing electron concentration the gate potential becomes screened, thereby preventing further carrier depletion in the channel.

On-state Performance

A number of factors can be drawn on to account for the significant improvement in drive current and transconductance of devices on c805 compared to c802. Part of the reason for the improved on-state performance is due to the better gate control of c805. This is attributed to the smaller CET arising from the inverted device design of the c805 wafer, for which the gate-to-channel distance is reduced by 5 nm over that of c802 wafer. The resulting improvement in CET is quantified to be 1.57 nm.

The transport properties of the grown wafers are also significantly different. The sheet resistance of the c802 wafer was previously determined to be 359.42 kΩ/sq. In comparison, the sheet resistance of the c805 wafer is 5× smaller, as shown in Section 7.2.3.2. This improvement in transport is likely to arise from the hole mobilities being higher in c805. Since the devices do not switch-off, an extraction of the effective mobility as a function of the carrier density could not be performed on c805 to enable a comparison to be drawn against c802. Instead numerical simulations of the hole population during device operation are undertaken to explain the suggested mobility improvement in c805. Illustrated in Figure 7.19 is the hole population at different regions of device operation based on 1D-PS simulations. As the device turns on, carriers populate towards the back interface of the channel (see Figure 7.19(a)). Therefore, mobility is likely to be limited by the effects of interface scattering from the back of the channel and ionised impurity scattering due to the dopants. With increasing gate bias the hole population in the channel increases, with the charge centroid still located at the back interface (see Figure 7.19(b)). As a result of the high hole accumulation the dopants are screened, therefore scattering due to ionised impurities is diminished, resulting in a peak in the hole mobility. Further increase in the gate bias results in the whole channel becoming populated with holes (see Figure 7.19(c)). Now the holes are subject to scattering from both the front and back interfaces, which is likely to cause a reduction in mobility. It is observed that a large fraction of holes still reside at the back interface where carrier mobility is highest, therefore the drop-off in mobility is inferred to be gradual. At a significantly high gate bias, apart from carrier population in the buried channel, a small fraction of holes begin to populate the parasitic quantum well (see Figure 7.19(d)). Interface and dielectric related scattering mechanisms would significantly degrade the mobility of these surface holes. However, the presence of a parasitic surface channel is not observed from the transfer characteristics shown in Figure 7.16(b).

The smaller sheet resistance of the c805 wafer also means that the contribution of

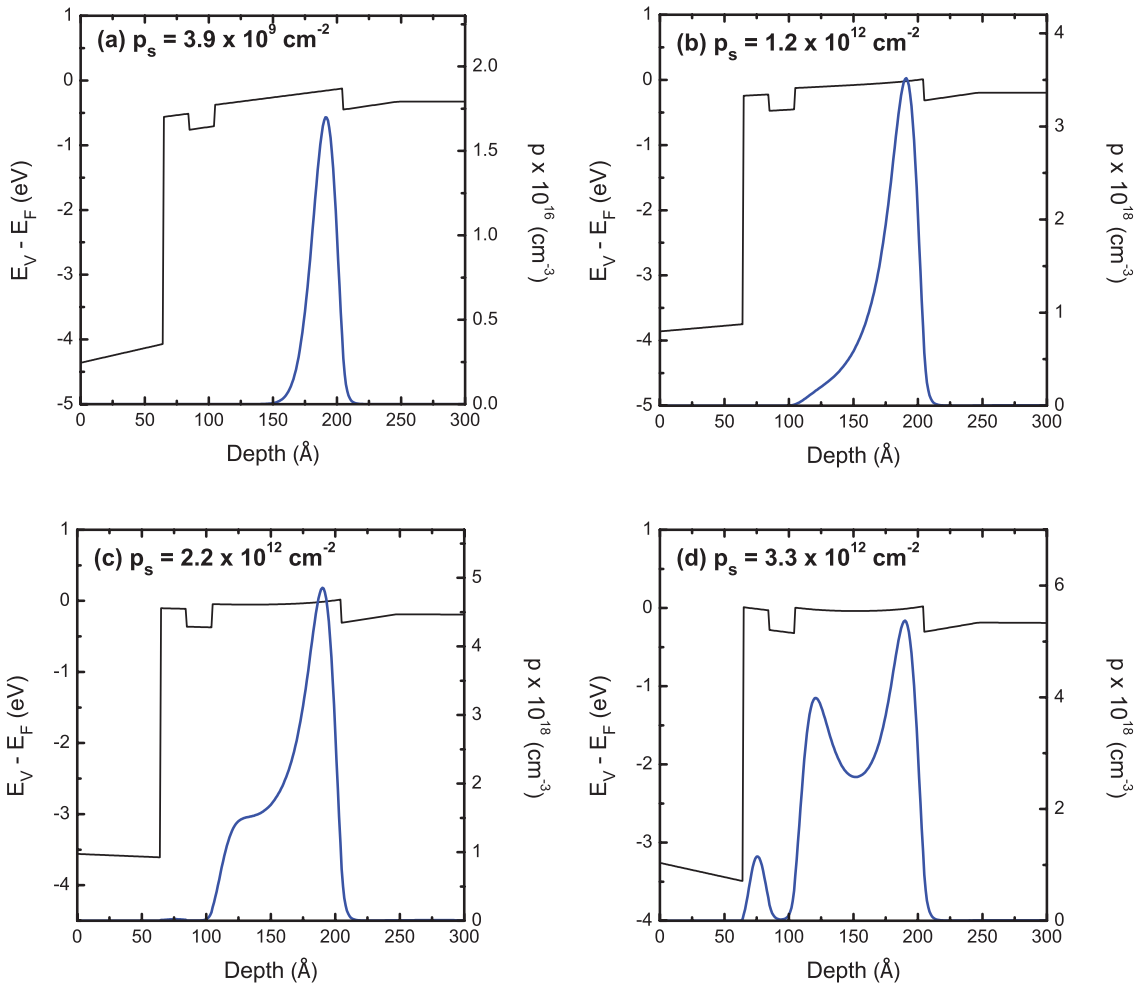


Figure 7.19: One-dimensional Poisson-Schrödinger simulations of hole populations of the c805 device.

access resistance to R_{sd} would be substantially lowered for the device. The contact resistance was determined to be 1.82 Ω.mm based on TLM measurements. Inserting the values of R_{c} and R_{sh} into Equation (6.4) results in a R_{sd} of 69.1 Ω.mm. This is a factor of 5 smaller than the series resistance of the c802 device. As a result, a larger fraction of the gate and drain biases are dropped across the intrinsic regions of the device, rather than being lost to parasitic series resistance. This is partly responsible for the higher drive current and transconductance of the c805 device, however access resistance still makes up approximately 90% of the total R_{sd}. Scaling the contact-to-gate spacing into the sub–μm regime would greatly reduce the contribution of R_{side} to R_{sd}. In the process significant improvements in the on-state performance is likely to be realised. This forms the motivation for the L_{side} gap scaling study presented in Section 7.2.3.2.

It was previously highlighted that D_{it} is likely to be at its lowest in the upper half of

the bandgap towards the conduction band. From Figure 7.18 it is observed that Fermi level moves through this energy range in the saturation regime of device operation. A smaller D_{it} distribution translates to better gate control of the channel, resulting in higher performance in the on-state.

7.2.3.2 Gate-to-Source/Drain Gap Dependence

The aforementioned analysis of the c805 device revealed that the on-state performance is being limited by the source/drain series resistance. Apart from the high sheet resistance, the large contact-to-gate spacing of 1 μm is the other key factor responsible for the high series resistance. Therefore, the downsizing of L_{side} gaps would minimise the impact of R_{sd} on device performance. In view of this, three gap sizes of 70 nm, 200 nm and 465 nm were chosen for investigation. The implementation of these scaled gaps in device fabrication required process development. Details of the developed process module can be found in Section 6.4. Devices featuring the scaled gaps were fabricated employing the process flow shown in Figure 6.22. In addition, devices featuring the standard 1 μm gap were also fabricated to allow for a comparison to be drawn against scaled gaps. Mask design for fabrication encompassed devices of all four gaps. This ensured that all devices were subject to identical processing, thereby eliminating process variations to device results. This was particularly important given the run-to-run variability of Au/Zn/Au ohmic metallisation which could have a large bearing on the resultant analysis. Although devices of different gate lengths (300 nm, 1 μm, and 20 μm) were fabricated for each gap size, the following discussion will be restricted to 1 μm–L_{g} devices.

The drain current and transconductance of devices at all four gaps are compared in Figure 7.20 at drain biases of -50 mV and -3.5 V. There is clear improvement in drain current and transconductance as L_{side} is scaled from 1 μm down to 70 nm. At $V_{\mathrm{d}} = -50$ mV and $V_{\mathrm{g}} = -3$ V, the 70 nm - L_{side} device shows a factor of 5.2 enhancement in drain current compared to the 1 μm–L_{side} device. The transconductance is also improved by a factor of 4.4. At the drain bias of -3.5 V and gate bias of -3 V, the drain current is enhanced by 3.6× and the transconductance presents with 4.1× enhancement as the gap is scaled from 1 μm to 70 nm. A maximum $I_{\mathrm{d,sat}}$ of 72.4 mA/mm and a peak g_{m} of 25.8 mS/mm are obtained from the transfer characteristics of the 70 nm gap at $V_{\mathrm{d}} = -3.5$ V.

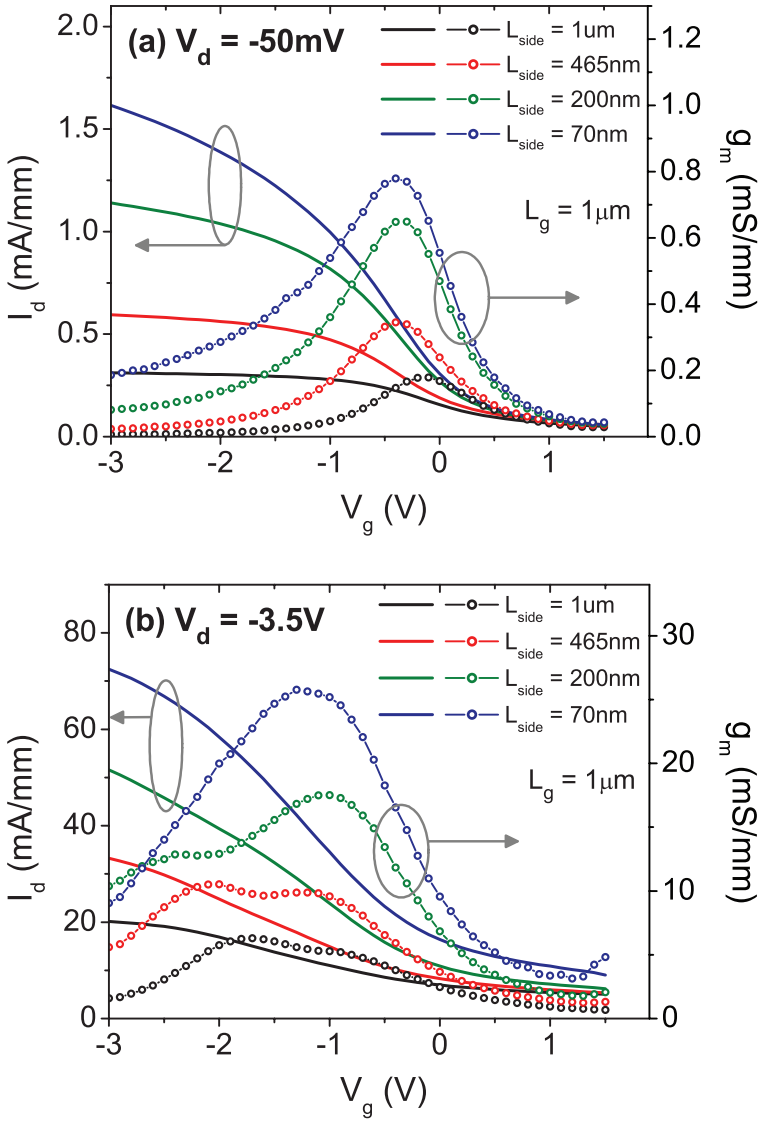


Figure 7.20: Transfer characteristics of c805 devices with $L_g = 1\,\mu$m and gate-to-source/drain gaps, L_{side}, of 1 μm, 465 nm, 200 nm and 70 nm at (a) $V_d = -50$ mV and (b) $V_d = -3.5$ V.

Illustrated in Figure 7.21 are the output characteristics of the devices with each of the four L_{side} gaps. From the output conductance it is noted that the effect of impact ionisation is enhanced for the devices with smaller gaps. This is a consequence of the higher drain field arising from the minimisation of potential loss to the series resistance. Impact ionisation could be partly contributing to the increase in drain current for the smaller gap devices. The leakage floor is found to be fairly similar between devices featuring 200 nm, 465 nm and 1 μm L_{side} gaps. However the leakage floor of the 70 nm - L_{side} device is 1.5× higher. The on-resistance extracted from the I_d–V_d plots at $V_g = -3$ V

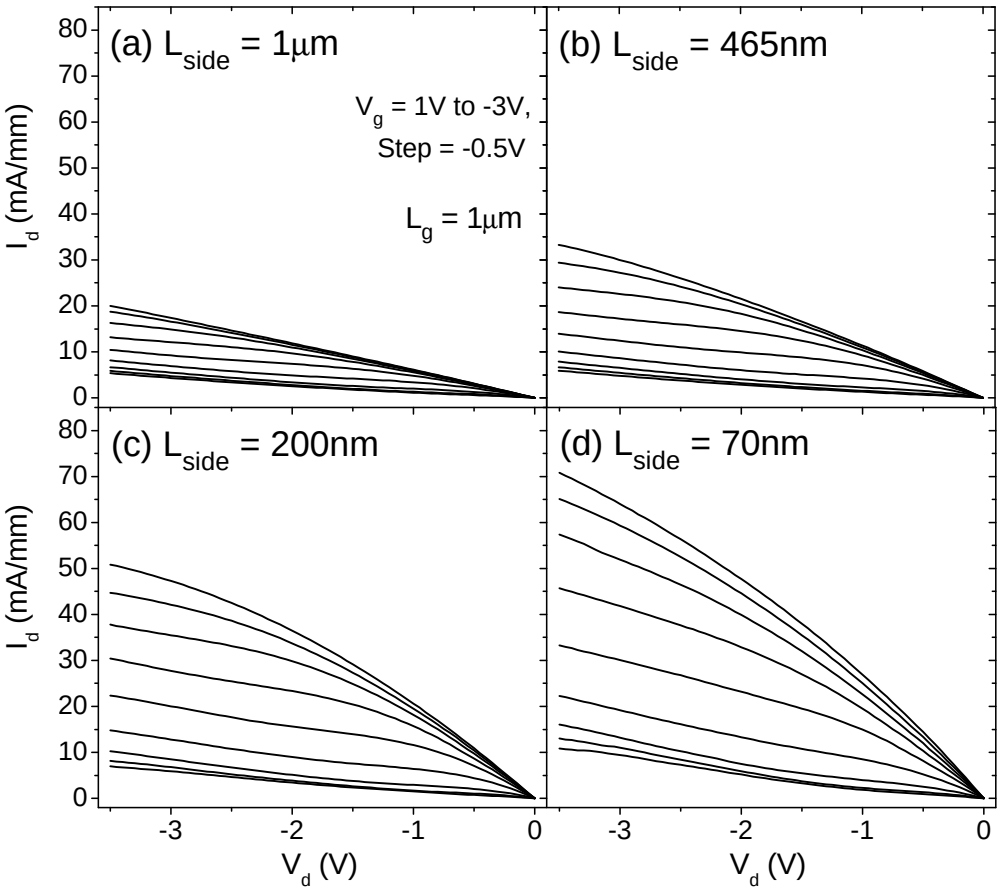


Figure 7.21: Output characteristics of the c805 devices with $L_g = 1\,\mu$m and gate-to-source/drain gaps, L_{side}, of (a) $1\,\mu$m, (b) 465 nm, (c) 200 nm and (d) 70 nm.

and $V_d = -50$ mV decreased from 160 k$\Omega.\mu$m for the $1\,\mu$m–L_{side} device to 31 k$\Omega.\mu$m for the 70 nm - L_{side} device.

The sheet resistance and contact resistance of the c805 device was extracted using a combination of device data obtained at different gate lengths and at different contact-to-gate spacings. The extraction procedure is fairly straightforward. In the first instance, on-resistance as a function of gate length is plotted for each of the L_{side} gaps. This is illustrated in Figure 7.22(a). The total series resistance corresponding to each of the L_{side} gaps is evaluated from the intercept of the fitted line with the y-axis. Given that the contact resistance between the devices should be identical (or at least fairly similar), this would act as a constant in Equation (6.4). As a result, R_{sd} can be plotted as a function of the total L_{side} comprising the gap spacings at both the source and drain side of the transistor, as shown in Figure 7.22(b). The total contact resistance is determined from the intercept of the linear fit with the y-axis, which results in a R_c of 1.13 Ω.mm. From the gradient of the fitted line the sheet resistance is determined to be 67.3 kΩ/sq.

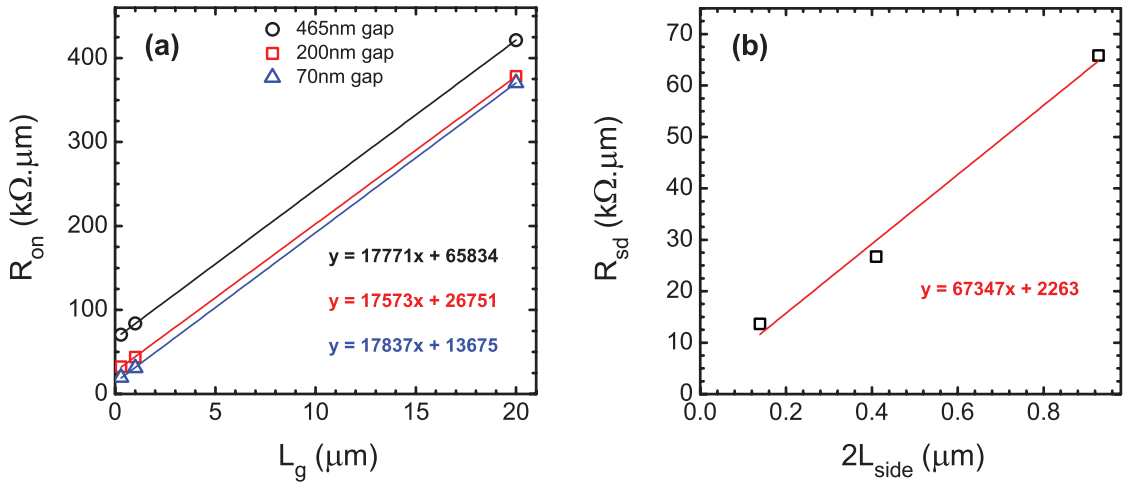


Figure 7.22: Plot of (a) R_{on} as a function of gate length for L_{side} gaps of 70 nm, 200 nm and 465 nm and (b) extracted R_{sd} as function of gap spacings.

7.2.4 Summary

InGaAs-channel MOSFETs based around a 10 nm thick, 75% indium-content channel were fabricated in this work. The device layers were based on the buried-channel design concept. Two wafers were grown: c802 was based on the non-inverted design while the inverted design was adopted in c805. $2\,\mu$m–L_{g} transistors at a fixed L_{side} of $1\,\mu$m were realised using an identical process flow. The technique of arsenic capping/de-capping was employed for interface passivation, while Al_2O_3 was used as the gate dielectric.

The c802 device showed good current modulation, saturation behaviour and switch-off characteristics, however several limitations were evident in the device characteristics. The device presented with a positive threshold voltage of 0.189 V, a modest saturation drain current of 1.14 mA/mm, a large subthreshold swing of 300 mV/dec and a small on-off current ratio of 74. Large parasitic source/drain resistance (732 Ω.mm), a large number of interface traps in the lower half of the bandgap and the presence of a parasitic surface channel further degrade the performance of the device. Underlying causes for the non-ideal aspects of device performance in both the on and off-state were presented and argued.

The c805 device also demonstrated good control of drain current by the gate. The on-state performance was found to be substantially improved compared to the c802 device under the same gate/drain bias conditions ($I_{\text{d,sat}} = 12.82$ mA/mm, $g_{\text{m}} = 5.85$ mS/mm). On-resistance was also found to be a factor of 5.6 lower. Smaller CET of the inverted design, better material transport of the starting wafer, smaller R_{sd} as well lower D_{it} levels were all found to contribute to the improved on-state performance. However the

off-state characteristics were degraded for c805. The device suffered from substantial subthreshold leakage current, which prevented switch-off to be achieved. Chief among the arguments put forth to explain the high off-state current were buffer leakage and Fermi level pinning in the region of subthreshold operation. In addition, ways of advancing the overall performance of c805 device were discussed.

Gate-to-source/drain contact gap scaling was also examined on c805. Scaling the gap from 1 μm down to 70 nm, with the gate length fixed at 1 μm, produced a factor of 3.6 enhancement in drain current and a factor of 4.1 improvement in device transconductance. A maximum $I_{\mathrm{d,sat}}$ of 72.4 mA/mm and a peak g_{m} of 25.8 mS/mm were obtained at the smallest gap.

7.3 InGaSb-Channel Device Performance

7.3.1 Device Specification and Processing

Figure 7.23 depicts the layer structure of the InGaSb-channel device wafers. Both wafers are based on the inverted design concept. Similar to the InGaAs-channel wafers, a GaSb cap was used as an in-situ passivant to suppress the oxidation of the AlSb barrier layer and improve the quality of the high-k/III-V interface [93, 207, 259]. A notable difference between the InGaAs and InGaSb-channel wafers though is the thickness of the cap. Inclusion of a thicker cap layer in InGaSb-channel wafers caters for the GaSb etch resulting from the ex-situ surface treatment prior to gate dielectric deposition (see Section 6.5.3.2). AFM results indicated approximately 5.2 nm of p-GaSb is etched following 1% $(NH_4)_2S$ surface treatment. Although a 8 nm thick cap should ensure a GaSb layer of at least 2 nm in thickness remains after the treatment, as a precautionary measure the thickness of the grown cap layer was increased to 10 nm. The two wafers are differentiated by the thickness and material composition of the channel. A 10 nm thick, 30% In-content channel (1.23% strain) is employed in wafer 7-478, whereas wafer 7-479 features a thinner channel of 7.5 nm in thickness and a higher indium composition of 40% (1.8% strain).

An ex-situ $(NH_4)_2S$ surface treatment was employed on the air-exposed III-V epi wafers to passivate the interface prior to gate dielectric deposition. The sulphidation process was preceded by a degrease step comprising of a sequential rinse of the wafers for 1 min each in acetone, methanol and isopropanol. This was followed by a 10 min immersion in $(NH_4)_2S$ diluted to 1% concentration in deionised H_2O, with the aqueous solution held at room temperature ($\sim$295 K). These conditions were determined to be optimum for GaSb surfaces based on the interface passivation study presented in Section 6.5. The

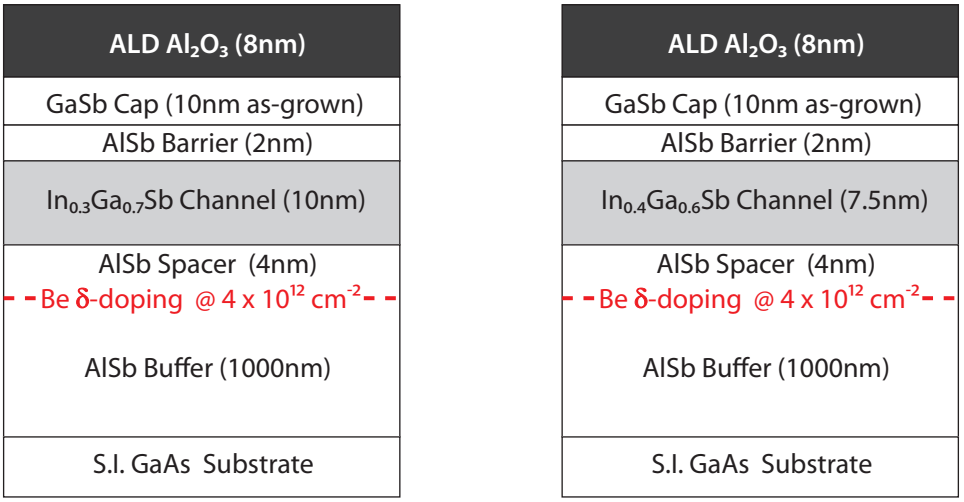


Figure 7.23: Layer structure of InGaSb-channel device wafers based on the implant-free, flatband MOSFET architecture incorporating the inverted design concept and featuring a biaxial compressively strained channel: (a) 7-478 with a 10 nm thick $In_{0.3}Ga_{0.7}Sb$ channel (1.23% strain) and (b) 7-479 with a 7.5 nm thick $In_{0.4}Ga_{0.6}Sb$ channel (1.8% strain).

treated samples were then loaded into the ALD tool within as short a lag time as possible from removal from the solution (∼4 min), in order to keep ambient contamination and regrowth of native oxides to a minimum [26]. A 8 nm-thick (nominal) Al_2O_3 gate dielectric film was formed by ALD using alternating pulses of TMA and H_2O precursors at 300°C, in a TMA-first process. Based on TEM imaging, the physical thickness of the Al_2O_3 film was determined to be 8 nm (see Figure 6.26(a)).

Processing for device realisation was identical between the two wafers. Transistors were fabricated using a GF process flow (see Figure 6.3), consisting of a combined gate/marker level and a separate ohmic level. A detailed listing of this fabrication process is included in Appendix A.4. All devices feature a 1 μm-L_{side} and 100 μm-W_g. While transistors of several gate lengths (500 nm - 20 μm) were fabricated, the following discussion will be primarily focussed on the 2 μm-L_g devices. The measured results from both wafers are presented and reviewed in a comparative manner.

7.3.2 Results and Discussion: 7-478 vs. 7-479

Shown in Figure 7.24 are the typical DC output and transfer characteristics of MOSFETs based around a 10 nm thick, $In_{0.3}Ga_{0.7}Sb$-channel (7-478) and a 7.5 nm thick, $In_{0.4}Ga_{0.6}Sb$-channel (7-479). The measured devices feature a gate length of 2 μm and a gate-to-source/drain separation of 1 μm.

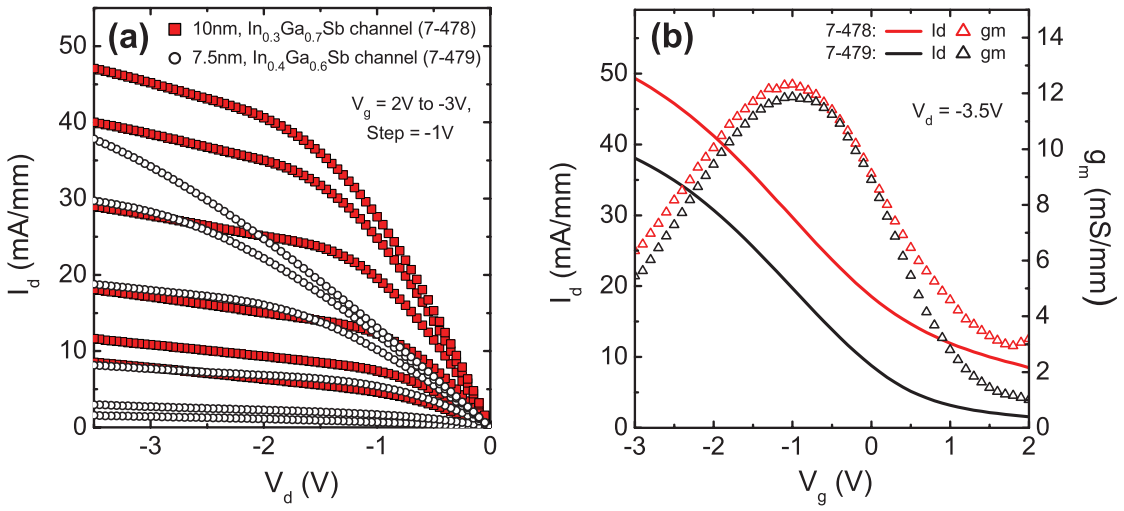


Figure 7.24: Measured (a) output and (b) transfer characteristics of the 2 μm - L_g, 1 μm - L_{side} p-MOSFETs fabricated on 7-478 and 7-479 device wafers.

From initial inspection it is seen that the devices exhibit well-controlled current modulation and good saturation behaviour (see Figure 7.24(a)). A maximum saturation drain curent of 49.43 mA/mm and 38.04 mA/mm are obtained for the 7-478 and 7-479 devices respectively, under a gate bias of -3 V and a drain bias of -3.5 V. The transconductance values of both devices, evaluated at a drain bias of -3.5 V, are very similar (see Figure 7.24(b)). A g_m of 12.32 mS/mm is obtained for the device on 7-478 while the 7-479 device presents with a g_m of 11.87 mS/mm. On the other hand, similar to the off-state performance of the transistor fabricated on c805 (InGaAs-channel wafer based on the inverted design), both InGaSb-channel devices do not switch-off. However the off-state leakage as observed from Figure 7.24(b) is higher for 7-478 compared to 7-479. The off-state current of the 7-478 device is 8.49 mA/mm at $V_d = 2$ V and $V_g = -3.5$ V. Under similar gate and drain bias conditions, the device on 7-479 exhibits an I_{off} that is 5.4× smaller. The differences in performance between the two devices are discussed in the remainder of this section, which is organised into three parts. In the first two parts, the off-state and on-state performance are presented. Following this the performance metrics

of InGaSb-channel devices realised in this work are compared against transistors reported in literature.

7.3.2.1 Off-State Performance

An analysis of the poor switch-off characteristics of c805, the InGaAs-channel device, was previously presented (see Section 7.2.3.1). It is worth noting that a similar analysis also applies to the InGaSb-channel devices to account for the degraded off-state performance, therefore these arguments will not repeated here. Rather, the following discussion is intended to elucidate the underlying reasons giving rise to the differing off-state performance between 7-478 and 7-479 devices.

Shown in Figure 7.25 are the semi-logarithmic plots of the I_d–V_g and I_g–V_g curves of the 7-478 and 7-479 devices. It is seen that the gate leakage current is several orders of magnitude lower than the drain current at all drain biases in the investigated gate bias range. Therefore, gate leakage plays no part towards the observed high subthreshold drain currents of the InGaSb-channel devices.

With the exception of the thickness and material composition of the channel, the 7-478 and 7-479 devices are identical in all other respects. This includes interface passivation and processing for transistor fabrication. Based on this, the difference in the off-state leakage currents between the devices is likely to arise from the Fermi level moving through differing energy levels at the interface of Al_2O_3/GaSb. Illustrated in Figure 7.26 are the relevant

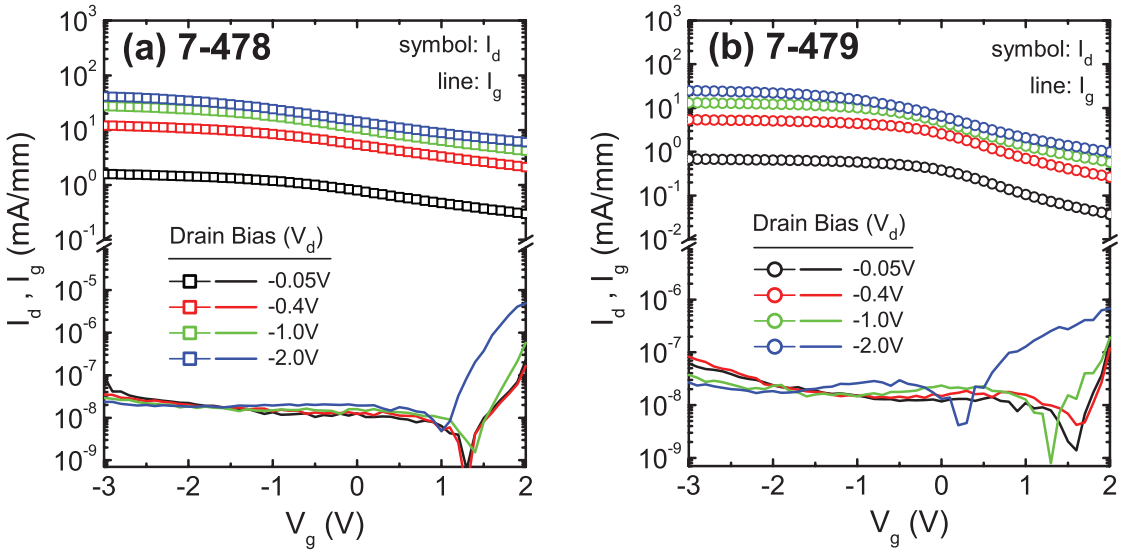


Figure 7.25: Semi-logarithmic plot of the measured I_d–V_g and I_g–V_g characteristics of $2\,\mu$m - L_g, $1\,\mu$m - L_{side} p-MOSFETs realised on (a) 7-478 and (b) 7-479 device wafers.

trap energies associated with the 7-478 and 7-479 transistor operation calculated based on 1D-PS simulations. In the case of 7-478, the Fermi level is already at 0.18 eV above the conduction band edge of GaSb (cap layer) as the device enters subthreshold operation. On the other hand, the Fermi level is located closer to the conduction band edge in the subthreshold region of 7-479 device operation. Based on the sulphide passivation study of Al_2O_3/GaSb interfaces (see Section 6.5), it was concluded that Fermi level movement is heavily obstructed by a large density of interface traps in the upper half of the bandgap towards the conduction band. Although trap densities in this region of the bandgap were not determined, based on the D_{it} profile shown in Figure 6.30 it is inferred that the trend of ascending distribution in the lower half of the bandgap carries into the upper half and into the conduction band. It is highly likely that D_{it} numbers in the high 10^{13} or even in the $10^{14}\,cm^{-2}eV^{-1}$ exist close to and into the conduction band. This implies FLP occurs at trap energies located close to, or just a little into, the conduction band. For sake of argument, the pinning position is assumed to be at $E_{tr} = 0.05\,eV$. The channel carrier density corresponding to this trap energy level is demarcated in Figure 7.26. It is noted that at the onset of FLP, the channel carrier density is lower for 7-479 in comparison to 7-478. This means the gate electrode is able to deplete more of the carriers in the 7-479 device before the potential is screened from the channel. This would account for the subthreshold drain current being smaller for 7-479 compared to 7-478. Alternatively, if the pinning is assumed to occur farther into the conduction band, for instance at 0.18 eV, then it is possible the channel is completely depleted in the 7-479 device. In this case

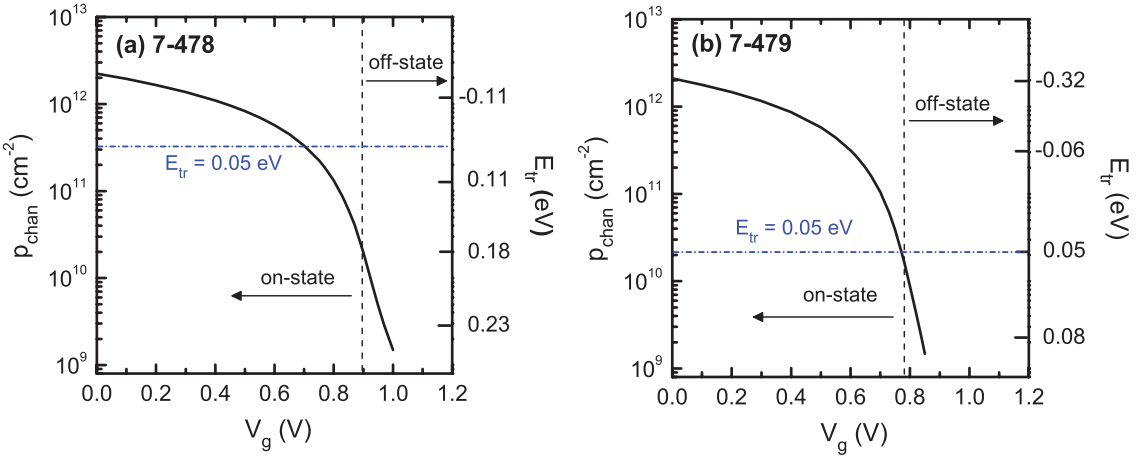


Figure 7.26: Plot of channel carrier density and trap energy as a function of gate bias for (a) 7-478 and (b) 7-479 device wafers based on Poisson-Schrödinger simulations. Trap energy (E_{tr}) is defined as zero at the conduction band edge and negative within the bandgap.

buffer leakage could be preventing complete switch-off in 7-479. However, due to FLP the channel of the 7-478 device would still be populated with carriers. It is postulated that the combination of FLP and leakage through the buffer is responsible for the observed off-state leakage currents in InGaSb-channel devices. Moreover, undoped InGaSb channels have been shown to possess residual p-type concentration [76, 92] and could possibly account for the poor switch-off characteristics.

7.3.2.2 On-State Performance

InGaSb-channel devices demonstrate a clear improvement over the c805 device (InGaAs-channel wafer based on the inverted design) with respect to saturation drain current and maximum transconductance. For $V_{\mathrm{d}} = -2\,\mathrm{V}$ and $V_{\mathrm{g}} = -3\,\mathrm{V}$, the drain current of 7-478 and 7-479 are 3× and 2× higher respectively compared to c805. In terms of transconductance, a factor of 2 and 1.6 enhancement is obtained for 7-478 and 7-479 in relation to c805. This is to be expected given that hole mobilities in antimony-based quantum well channels are higher to start with in comparison to InGaAs-channels, even without strain induction. The incorporation of strain in the channel would only act to further improve the mobility of holes. In addition to this, the InGaSb-channel devices of this work benefit from a higher valence band offset of 0.43 eV at the quantum well interfaces. On the other hand, the band offset of c805 is only 0.3 eV. The larger offsets of the InGaSb-channel devices provide for improved carrier confinement in the QW. The higher mobilities and enhanced carrier confinement enable better on-state performance to be attained in InGaSb-channel devices.

Lower on-resistance values were also achieved in InGaSb-channel devices. From the I_{d}–V_{d} plot at $V_{\mathrm{d}} = -50\,\mathrm{mV}$ and $V_{\mathrm{g}} = -3\,\mathrm{V}$, the on-resistance of the 7-478 device was extracted to be $31.7\,\mathrm{k\Omega.\mu m}$. In a similar vein, the on-resistance of 7-479 was determined to be $73.5\,\mathrm{k\Omega.\mu m}$. For the purpose of quantifying the series resistance associated with the InGaSb devices, device on-resistance as a function of gate length is plotted in Figure 7.27 for 7-478 and 7-479. From this R_{sd} is evaluated to be $28.7\,\Omega$.mm for 7-478 and $55.9\,\Omega$.mm for 7-479. The sheet resistance values of the device wafers were determined from Hall measurements for which the results are presented in Table 7.1. Using the known values of R_{sd} and R_{sh}, the contact resistance of the devices were calculated from Equation (6.4). The contact resistance of the 7-478 device was $2\,\Omega$.mm, while that of the 7-479 device was $5.32\,\Omega$.mm.

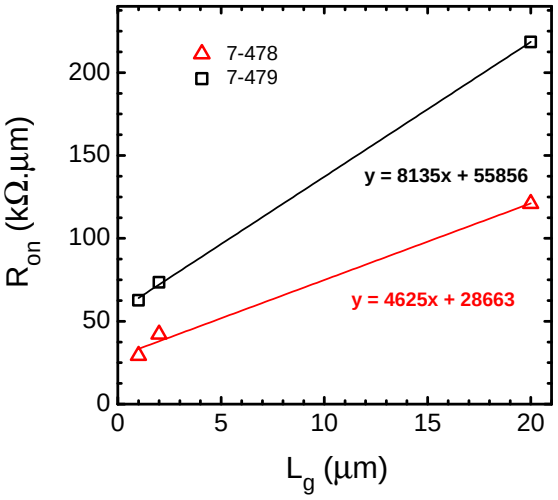


Figure 7.27: R_{on} as a function of gate length for the 7-478 and 7-479 devices, measured at $V_{\mathrm{g}} = -3\,\mathrm{V}$ and $V_{\mathrm{d}} = -50\,\mathrm{mV}$.

A closer inspection of the device characteristics, coupled with the Hall data, reveal a somewhat unexpected discrepancy in the on-state performance between the InGaSb-channel transistors. Contrary to the device with the thinner and higher indium content InGaSb-channel (7-479) presenting with higher drive current and transconductance, better on-state performance is in fact derived from the 7-478 device comprising a thicker and less strained channel. This should not occur given that higher mobility and better carrier confinement is to be expected of 7-479. Based on the Hall measurement though it seems the mobility of 7-479 is smaller than that of the 7-478 wafer by a factor of two, while the carrier densities only differ by a margin of 15%. This suggests that carrier scattering from the back and front interfaces of the channel is dominant for devices employing AlSb layers as a barrier, spacer and/or buffer. With decreasing channel thickness the holes populating the channel are likely subject to a higher degree of interface scattering, which in turn would limit the peak channel mobility that is reached. There are two ways this mobility-limiting mechanism can be overcome in the 7-479 layer design. The obvious way would be to relax the channel thickness. Alternatively the strain in the channel can be enhanced

Table 7.1: Van der Pauw data of the 7-478 and 7-479 device wafers.

Wafer	R_{sh} (kΩ/sq)	μ_{h} (cm^2/Vs)	p_{s} (cm^{-2})
7-478	12.3	289	1.76×10^{12}
7-479	22.7	136	2.03×10^{12}

by adopting an even higher indium composition for InGaSb, while still maintaining the channel thickness. A combination of relaxed channel thickness coupled with higher strain would greatly benefit the transport properties of 7-479. Such modifications would enable 7-479 to achieve mobilities higher than that of 7-478, thereby enhancing the resulting on-state device performance.

During the off-state discussion it was highlighted that the trap densities in the upper half of the bandgap are likely to be higher in contrast to the lower part of the bandgap. From Figure 7.26 it is noted that both InGaSb-channel devices appear to be operating in the upper part of the bandgap, it is therefore possible that the efficiency of the gate control to modulate the channel charge is being affected. Shifting the device operation into the lower half of the bandgap, where D_{it} reaches values into the mid $10^{12}\,\mathrm{cm}^{-2}\mathrm{eV}^{-1}$ (see Figure 6.30), would improve gate control. This would enhance the on-state performance significantly. The off-state performance would also see improvement since it is moved farther away from the FLP position, which in turn would enable the devices to be turned off. This can be achieved with the adoption of lower work function gate metal. Device performance would further benefit from EOT and L_{side} gap scaling.

7.3.2.3 Performance Benchmark

As a result of the focus within the III-V community being centred around n-channel devices, there has only been a limited number of reports on III p-channel MOSFETs. Regardless, it is worth benchmarking the performance of InGaSb-channel devices of this work against reported literature.

A summary of the state-of-the-art device results on $\mathrm{In}_x\mathrm{Ga}_{1-x}\mathrm{Sb}(100)$-based p-channel MOSFETs, comparing $I_{\mathrm{d,sat}}$, g_{m}, SS, $I_{\mathrm{on}}/I_{\mathrm{off}}$ and peak field-effect hole mobility ($\mu_{\mathrm{h,FE}}$), is given in Table 7.2. Due to differences in device dimensions and architecture, and choice and thickness of gate dielectric, among other differences, a like-for-like comparison between the devices is challenging. The influence of device dimensions however is excluded by normalising $I_{\mathrm{d,sat}}$ and g_{m} with respect to the gate lengths and widths specific to each device. The corresponding driving voltages, as reported in each reference, are further listed below Table 7.2. The following comparison is limited to the on-state performance metrics of $I_{\mathrm{d,sat}}$ and g_{m} as the other metrics could not be evaluated for InGaSb-channel devices of this work, due to incomplete device switch-off.

Table 7.2: Summary of representative results on $In_xGa_{1-x}Sb(100)$-based p-channel MOSFETs.

Device architecture	Channel	Source/Drain technology	Dielectrics	EOT (nm)	$I_{d,sat} \times L_g$ (mA/mm).μm	$g_m \times L_g$ (mS/mm).μm	SS (mV/dec)	I_{on}/I_{off}	$\mu_{h,FE}$ (cm^2/Vs)	Ref.
Inversion type	GaSb	SA implant	Al_2O_3	3.9	52.5	19.5	~600	265	200	[84] [a]
Inversion type	GaSb	SA implant	Al_2O_3	4.88	22.5	···	~260	$>10^3$	290	[85] [b]
Inversion type	GaSb	SA implant	Al_2O_3	4.88	24	···	250	$>10^3$	160	[86] [c]
Inversion type	GaSb	SA implant	Al_2O_3/RT-Y_2O_3	2.47	60	50	~400	4	80	[87] [d]
Inversion type	GaSb	SA implant	Al_2O_3/HT-Y_2O_3	2.47	130	90	147	10^3	200	[87] [d]
Inversion type	GaSb	SA nickelicide	Al_2O_3	2.44	20	···	~400	~10	25	[88] [e]
Inversion type	GaSb	SA nickelicide	Al_2O_3	3.9	~3	···	250	~50	~270	[89] [f]
Inversion type	GaSb	SA nickelicide	Al_2O_3	4.88	4.25	···	~320	~50	~11	[90] [g]
Surface-channel QW	$In_{0.2}Ga_{0.8}Sb$	SA nickelicide	Al_2O_3	3.9	210	145	140	$>10^4$	~510	[89] [h]
Surface-channel QW	$In_{0.36}Ga_{0.64}Sb$	SA implant	HfO_2	1.95	69	39	···	···	···	[91] [i]
Surface-channel QW	$In_{0.35}Ga_{0.65}Sb$	SA implant	Al_2O_3	4.88	360	470	120	$>10^4$	740	[92, 93] [j]
Buried-channel QW	$In_{0.35}Ga_{0.65}Sb$	SA implant	Al_2O_3	4.88	···	700	125	$>10^4$	910	[92, 93] [j]
Buried-channel QW	$In_{0.3}Ga_{0.7}Sb$	LA Au/Zn/Au	Al_2O_3	3.9	98.9	24.6	···	···	···	This work
Buried-channel QW	$In_{0.4}Ga_{0.6}Sb$	LA Au/Zn/Au	Al_2O_3	3.9	76.1	23.8	···	···	···	This work

[a] $V_g = -4\,V$ and $V_d = -3\,V$
[b] $V_g = -3.75\,V$ and $V_d = -2\,V$
[c] $V_g = -3\,V$ and $V_d = -3\,V$
[d] $V_g = -1.5\,V$ and $V_d = -1.5\,V$
[e] $V_g = -3\,V$ and $V_d = -1\,V$
[f] $V_g = -1.5\,V$ and $V_d = -0.5\,V$
[g] $V_g = -2\,V$ and $V_d = -1\,V$
[h] $V_g = -2.5\,V$ and $V_d = -2\,V$
[i] $V_g = -3\,V$ and $V_d = -2\,V$
[j] $V_g = -2\,V$ and $V_d = -2\,V$

It is noted that the drain currents of antimonide devices realised in this work are among the highest reported values. In contrast to this the transconductance values are among the lowest reported values. The primary reason for this is the large gate-to-source/drain gaps in devices of this work, which causes a dramatic reduction in the extrinsic transconductance numbers. On the other hand, the devices reported in literature employ self-aligned implant or nickelicide for source/drain junction formation. A factor of four improvement in transconductance was demonstrated for InGaAs-channel devices, realised in this work, by scaling gap spacings from 1 μm to 70 nm (see Section 7.2.3.2). A similar gap scaling exercise for InGaSb-channel devices would result in higher, and therefore more competitive, g_{m} values, as well as higher drain currents, to be attained.

7.3.3 Summary

InGaSb-channel MOSFETs based on the buried channel, inverted design concept were realised for investigation of device characteristics. Two wafers were grown in order to compare the effect of channel variations in terms of thickness and material composition. The 7-478 wafer featured a 10 nm thick, 30% indium-content channel, whereas a 7.5 nm thick, 40% indium-content channel was adopted in the 7-479 wafer. Both wafers were subject to an ex-situ ammonium sulphide treatment, the choice of interface passivation adopted for GaSb surfaces based on the interface engineering study undertaken in Chapter 6. Similar to InGaAs-channel devices, Al_2O_3 was employed as a gate dielectric in the InGaSb devices, but with a thickness of 8 nm. Devices were fabricated using a common process flow. Device analysis was restricted to the 2 μm–L_{g}, 1 μm–L_{side} MOSFETs.

Both InGaSb-channel devices demonstrated good gate modulation of the drain current. High drive currents and reasonable transconductance numbers were realised for 7-478 ($I_{\mathrm{d,sat}}$ = 49.4 mA/mm, g_{m} = 12.3 mS/mm) and 7-479 ($I_{\mathrm{d,sat}}$ = 38 mA/mm, g_{m} = 11.9 mS/mm) devices. The on-state characteristics of these devices were found to outperform that of c805 (InGaAs-channel device based on the inverted design). The improvement in performance for InGaSb-channel devices was attributed to the higher channel mobilities and better carrier confinement. Based on Hall measurements, it was found that InGaSb-channel wafers have a significantly lower sheet resistance in comparison to InGaAs-channel wafers. Modifications required to improve the on-state characteristics were also discussed. Suggestions included scaling the L_{side} gap and the EOT, as well as shifting device operation into the lower half of the bandgap where D_{it} levels are lower.

Similar to the off-state performance of c805, 7-478/7-479 devices could not be switched off. Although the subthreshold leakage current of 7-479 was 5.4× lower in comparison to

7-478. The possible causes that would account for the difference in off-state performance between the devices was pursued. An argument based on the FLP position in relation to the trap energies swept by the devices in subthreshold operation was put forward. The higher off-state leakage of 7-479 was rationalised on the basis that in subthreshold the device operates farther into the conduction band where pinning is expected to occur as opposed to 7-479 being on the fringes of the conduction band. As a result, the pinning of 7-478 occurs at a point where a substantial number of carrier reside in the channel, whereas more of the carriers would already have been depleted in 7-479 before it gets pinned. The section was concluded by benchmarking the performance of InGaSb-channel devices of this work against antimony-based devices reported in literature.

8

Conclusions and Future Work

8.1 Conclusions

With conventional Si CMOS scaling approaching its limits, innovations based on non-Si materials and new device architectures are required to facilitate improvements in transistor performance and extend the logic device roadmap. Future performance enhancements require a reduction in supply voltage to minimise power dissipation in logic circuits. However, this negatively impacts the drive current and switching speed of Si transistors. This can be mitigated by replacing the Si channel with alternate channel materials, in which charge carriers traverse at much higher velocities. In this regard III-V materials, owing to their superior electron transport properties, appear to be a promising n-channel candidate for post-Si CMOS.

Several key challenges, however, need to be addressed before III-Vs can be deployed in CMOS. These include a high quality dielectric/III-V gate stack, a low-resistance source/drain (S/D) technology, heterointegration onto a Si platform and a viable III-V p-MOSFET. This Thesis specifically addressed the development and demonstration of planar III-V p-MOSFETs, to complement the n-MOSFET, thereby enabling an all III-V CMOS technology to be realised. This work explored the application of InGaAs and InGaSb material systems as the channel, in conjunction with Al_2O_3/metal gate stacks, for p-MOSFET development based on the buried-channel flatband device architecture. The body of work undertaken comprised material development, process module development and integration into a robust fabrication flow for the demonstration of p-channel devices. The main findings and accomplishments of this Thesis are summarised below:

- The parameter space in the design of the device layer structure, based around the III-V channel/barrier material options of $In_{x\geq 0.53}Ga_{1-x}As/In_{0.52}Al_{0.48}As$ and

$In_{x\geq0.1}Ga_{1-x}Sb$/AlSb, was systematically explored to enhance channel transport, with emphasis on the hole mobility. The impact of the spacer and channel thickness, biaxial compressive strain, δ-doping density and doping strategy on channel transport were examined using pre-MOSFET structures, based on the flatband device architecture. Channel carrier density and mobility were determined from Hall measurements. Lattice-matched $In_{0.53}Ga_{0.47}As$ pre-MOSFETs demonstrated modest room temperature (RT) hole mobilities (60-70 cm^2/Vs). The modest values are a result of parallel conduction as determined from temperature-dependent Hall measurements. In comparison, $In_{0.85}Ga_{0.15}As$ pre-MOSFETs incorporating 2.1% strain exhibited 4.5-7× higher hole mobility, depending on the doping strategy. Dopant segregation giving rise to mobility degradation in inverted structures was suppressed by reducing the growth temperature from 450 °C to 400 °C, as confirmed by SIMS analysis. A key achievement is the mobility of 433 cm^2/Vs obtained for the non-inverted $In_{0.85}Ga_{0.15}As$ pre-MOSFETs (2.1% strain), which is the highest RT hole mobility of any InGaAs QW channel reported to date. The doping strategy and channel strain, with all other design parameters being equal, were studied for $In_{x\geq0.1}Ga_{1-x}Sb$ pre-MOSFETs. $In_{0.3}Ga_{0.7}Sb$ pre-MOSFETs featuring 1.21% strain demonstrated hole mobilities of 1096 cm^2/Vs (non-inverted) and 818 cm^2/Vs (inverted) compared to the mobility of 738 cm^2/Vs for the lattice-matched $In_{0.1}Ga_{0.9}Sb$ structure. However, lower channel carrier densities were obtained for strained pre-MOSFETs owing to the increase in the energy of valence band states, resulting in reduced charge transfer. The sheet resistance (R_{sh}) of these antimonide-based pre-MOSFETs are competitive with reported literature and only fall short of the results reported on the GaSb/AlAsSb material option which benefits form enhanced charge transfer due to the large valence band offset and double-sided doping.

- During the course of this work, a number of process modules that comprise the overall device flow were developed and experimentally validated as isolated entities prior to device integration. The first of these modules is ohmic contacts. It is noted that in this work ohmic contacts were developed purely to enable transistor realisation and not as a technological solution, advancement or integration. The implant-free nature of the flatband devices made the development and optimisation of contacts challenging. Thermally annealed Au/Zn/Au metallisation was investigated as a contact solution, with a 5 nm thick lower Au metal serving as a nucleation layer within the contact scheme. The contacts were developed on pre-MOSFETs based around InGaAs and InGaSb channels, capped with InGaAs and GaSb layers respectively.

The performance of contacts was evaluated from TLM structures. A series of experiments were conducted to investigate the effects of the Zn and upper Au layer metallisation thickness, a diffusion barrier in the contact scheme and de-oxidation conditions prior to metallisation on the resulting contact performance. A 1 HCl :4 H_2O for 30 s de-oxidation procedure prior to Au(5 nm)/Zn(40 nm)/Au(150 nm) metallisation annealed at 380°C for 120 s resulted in the lowest contact resistance (R_c) of 10.77±0.46 Ω.mm for contacts on InGaAs pre-MOSFETs. The best performing contacts on InGaSb pre-MOSFETs had a R_c of 1.862±0.35 Ω.mm, obtained for Au(5 nm)/Zn(40 nm)/Au(100 nm) metallisation annealed at 340°C for 120s, with a de-oxidation condition of 1 HCl : 2 H_2O, 60 s. A more concentrated HCl solution, in conjunction with a longer de-oxidation time, was required for contacts on InGaSb pre-MOSFETs as the de-oxidation procedure of 1 HCl : 4 H_2O for 30 s yielded poor and unreliable contact performance (R_c = 4.808±5.35 Ω.mm). One issue with this contact scheme is the possible run-to-run variation in contact performance between identically processed samples, likely arising from the variable deposition rate of Zn.

- In addition to optimising contacts for lower R_c, the access resistance, R_{side}, needed to be minimised to mitigate the effects of parasitic S/D series resistance on device performance. One contribution to R_{side} is the gap spacing, L_{side}, between the gate and S/D contacts, with smaller gaps providing lower R_{side}. Due to lithographic constraints, L_{side} was limited to 1 μm in the fabrication process previously developed in the group for n-MOSFET realisation, and also adopted for p-channel devices of this work. A fabrication process for deep submicron L_{side} scaling was therefore examined in this work. Self-aligned (SA) and lithographically-aligned (LA) techniques were reviewed to arrive at a methodology that was process compatible with III-V p-MOSFETs of this work. A LA double ohmic patterning approach, in which source and drain contacts are defined concurrently over two levels of lithography, was chosen for development. In this approach, stripe contacts (2 μm×100 μm) are first aligned to the gate and patterned. Submicron gaps are realised from limiting the lithography to well within the electron back scatter range, thereby alleviating proximity effect and its impact on resist exposure. Subsequently large contacts (150 μm×100 μm), required for measurements using wafer probes, are patterned to overlap the stripe contacts for electrical continuity. To realise S/D contacts that are ohmic across both the stripe and large contacts annealing is done post large contact patterning. The developed process yielded a variety of gap spacings from sub-100 nm up to 1 μm, with a key highlight being the ability to realise gaps as small

as 20 nm. Smaller gaps could not be realised due to issues with metal flagging, overlapping edges between the gate and stripe contacts, and metal lift-off. For device implementation, gaps smaller than 70 nm were deemed unsuitable based on the increased line edge roughness observed at these gap dimensions. Metal discontinuity, from erosion of substrate material, along the overlapping edge between the stripe and large contacts was revealed from the mechanical validation of the fabrication process on MOSFET material. Electrochemical etching resulting from the partial exposure of the stripe contact metal to the wet etching solution during the gate oxide etch, prior to large contact metallisation, was identified as the cause of substrate erosion. This was mitigated by removing the oxide etch step from the process flow. With large contacts serving as bondpads in the revised process the ohmic anneal was performed post stripe contact metallisation, without detriment to the resulting contact performance.

- The final module developed was the gate stack. As high-k/InGaAs gate stack modules, available within the Nonclassical CMOS Research Centre, were mature, no further development was needed. Focus, therefore, was placed on high-k integration on GaSb. Realising an unpinned high-k/GaSb interface, with a low defect density, is particularly challenging due to the presence of thick native oxides on GaSb. Among the reported studies, ammonium sulphide ($(NH_4)_2S$), a wet treatment that has shown a great deal of success for high-k/InGaAs interfaces, had only received little attention in high-k/GaSb systems. Thus, in this work, the effectiveness of $(NH_4)_2S$ treatments, in the range 1%-22%, was systematically investigated for improving the electrical properties of the high-k/GaSb interface. p-GaSb surfaces were treated in 1%, 5%, 10% and 22% $(NH_4)_2S$ solutions for 10 min at 295 K prior to ALD of Al_2O_3. The effects of the treatment were electrically assessed from frequency-dependent C–V measurements. While the 1% treated sample showed the largest capacitance swing, in conjunction with the smallest flatband voltage shift and stretch-out, a pinned Fermi level was obtained for the 22% treatment. A large interface trap density (D_{it}) response was evident in the upper half of the bandgap in all samples. The 1% treated sample had a U-shaped D_{it} distribution with a minimum of $4\times10^{12}\,cm^{-2}eV^{-1}$ at E_v+0.27 eV. The formation of interfacial layers for the 10% and 22% treatments, as revealed from TEM analysis, degraded the electrical properties reflected in the C–V characteristics. An extended study, comprising immersion times of 1, 5 and 10 min at each of the $(NH_4)_2S$ concentrations, was conducted to quantify the surface roughness as well as the etching nature of the

treatments on GaSb, based on AFM analysis. While no appreciable difference in roughness (1.37–1.45 nm) was noted for the 1% and 5% treatments, the roughness monotonically increased for the 10% and 22% treatments with increasing immersion times. Gate control in MOSFETs would be compromised from the increased surface roughness and interfacial layers at the higher $(NH_4)_2S$ concentrations. The pronounced chemical activity resulting from the higher sulphide concentrations and longer immersion times was also reflected in the thickness of GaSb etched. Knowledge of the treatment-dependent etch character was important in order to account for this during epi-layer growth. The 1% treatment, shown to be being most effective in this study, was adopted for the antimonide MOSFETs developed in this work.

- A number of III-V p-MOSFETs based on InGaAs and InGaSb-channels with the most promising hole transport, as determined from the pre-MOSFET investigation, and using the developed process modules were demonstrated in this work. The InGaAs-channel devices, non-inverted and inverted, employed a 10 nm thick $In_{0.75}Ga_{0.25}As$-channel and a 6.5 nm thick Al_2O_3 gate dielectric. Interface passivation was based on the technique of capping the epi-layers with As_2 during growth followed by As_2 decapping in the ALD tool prior to Al_2O_3 deposition. Non-inverted devices, with L_g=2 μm and L_{side}=1 μm, exhibited good current modulation, saturation behaviour and switch-off characteristics. Several aspects of performance however were non-ideal; depletion-mode operation (V_{th}=0.189 V), modest $I_{d,sat}$ (1.14 mA/mm), double peaked g_m (1.06 mS/mm), high SS (301 mV/dec), high R_{on} (845 kΩ.μm) and small I_{on}/I_{off} (74). A large access resistance of 718.84 Ω.mm, a high R_{sh} of 359.42 kΩ/sq, a high D_{it} in the lower half of the bandgap of device operation, a large CET of 5.93 nm and parasitic surface channel conduction were identified as causes for the non-ideal device characteristics. On the other hand, inverted devices, of identical device geometries to non-inverted devices, had substantially improved on-state performance. The metrics of $I_{d,sat}$, g_m and R_{on} were enhanced by 11×, 5.5× and 5.6× over that of non-inverted devices. This improvement is attributed to inverted devices having a CET that is 1.57 nm smaller, a 5× lower R_{sh}, a 5× smaller access resistance and a lower D_{it} in the on-state region of device operation. However inverted devices did not switch-off due to substantial subthreshold leakage, likely resulting from buffer leakage and Fermi level pinning. The benefit of scaling L_{side} from 1 μm down to 70 nm, for mitigating the impact of

access resistance on device performance, was demonstrated on 1 μm–L_g inverted devices. An $I_{d,sat}$ of 72.4 mA/mm and a g_m of 25.8 mS/mm, greater by a factor of 3.6 and 4.1 compared to 1 μm–L_{side} devices, were obtained for 70 nm–L_{side} devices. The InGaSb-channel devices, based on the inverted doping strategy, employed a 10 nm thick $In_{0.3}Ga_{0.7}Sb$-channel and a 7.5 nm thick $In_{0.4}Ga_{0.6}Sb$-channel. The ex-situ 1% $(NH_4)_2S$ treatment for 10 min at RT, developed in this work, was used for interface passivation prior to ALD of 8 nm-thick Al_2O_3. Both sets of devices, despite demonstrating well-controlled current modulation and good saturation behaviour, could not be switched off. In terms of the on-state characteristics, $In_{0.3}Ga_{0.7}Sb$-channel ($I_{d,sat}$ = 49.4 mA/mm, g_m = 12.3 mS/mm, R_{on} = 31.7 kΩ.μm) and $In_{0.4}Ga_{0.6}Sb$-channel ($I_{d,sat}$ = 38 mA/mm, g_m = 11.9 mS/mm, R_{on} = 73.5 kΩ.μm) devices outperformed the InGaAs-channel devices, for identical 2 μm–L_g, 1 μm–L_{side} transistor geometries. This was a result of the significantly lower R_{sh} of 12.3 kΩ/sq and 22.7 kΩ/sq in $In_{0.3}Ga_{0.7}Sb$-channel and $In_{0.4}Ga_{0.6}Sb$-channel devices, respectively. Despite the higher channel strain, the hole mobility of $In_{0.4}Ga_{0.6}Sb$-channel devices was a factor of 2.1 lower compared to $In_{0.3}Ga_{0.7}Sb$-channel devices. This indicated carrier scattering from channel/barrier interfaces, employing AlSb as the barrier, to be dominant in thinner channels. It is notable that the drain currents (normalised to L_g and W_g) of the InGaSb-channel devices of this work are among the highest values reported in literature. Based solely on the on-state performance, it can be argued that III-V p-MOSFETs based on InGaSb as opposed to InGaAs channels are more suited as the p-channel option for post-Si CMOS.

8.2 Future work

The focus of this work has been the development of materials and process modules for device integration, enabling the demonstration of III-V p-channel MOSFETs. Results based on the first iteration of devices indicate the performance potential of InGaSb as the p-channel solution in all III-V approach for post-Si CMOS. However, further material and process optimisation is required to tackle the shortcomings in the performance of InGaSb-channel devices. Some key areas for progression are listed below:

- **Channel transport:** As the body of work for layer optimisation was limited, an extended investigation of design parameter space should be carried out to fine tune

the layer structure for enhanced channel transport. To account for all mobility-limiting mechanisms, thus enabling for a more rigourous optimisation, the investigation should be based on MOSFET structures.

- **Gate stack:** The 1% $(NH_4)_2S$ treatment, shown to be most effective in this work, may not necessarily be the optimal treatment for Al_2O_3/GaSb interface passivation. Sulphide concentrations between 0% and 5%, for a variety of treatment times, should be further explored. In conjunction, FGA, well known for defect passivation in gate stacks, should be examined. Plasma-based treatments and low-temperature ALD, reported as being effective for high-k/GaSb interfaces, should form other avenues of study.

- **EOT:** For improvement in both the on- and off-state performance, the present EOT of 3.9 nm should be scaled to 1 nm or lower. Scaling the Al_2O_3 to $\leq$2 nm, to achieve this EOT target, is not practical due to the significant increase in gate leakage at these oxide thicknesses. A higher-k dielectric on its own or as a component of a bilayer gate stack, comprising Al_2O_3, should be investigated for EOT scaling.

- **Access resistance:** Based on the fabrication process for L_{side} scaling, developed in this work, devices with 70 nm–L_{side} should be realised to enhance $I_{d,sat}$ and g_m. For CMOS compatibility, SA fabrication processes should also be revisited. Due to the incompatibility of Si_3N_4 spacers, the fully SA process developed in the group to realise 70 nm gaps in III-V n-MOSFET was not used for p-channel device fabrication. A substitute spacer material should therefore be investigated.

A

Fabrication Process Flows

A.1 Hallbar Structure (Photolithography)

A.1.1 Ohmic

Clean Substrate	- 5 min 40 °C acetone, 5 min IPA, N_2 blow dry.
Spin Resist	- Microposit S1818, 4 krpm, 30 s.
Bake	- 90 °C oven, 15 min.
MA6 Exposure	- 3.8 s, hard contact.
Develop	- 75 s Microposit Developer:H_2O (1:1), 30 s H_2O rinse, N_2 blow dry.
Ash	- O_2 plasma, 40 W, 60 s.
De-oxidise	- 30 s HCl:H_2O (1:4), 30 s H_2O rinse, N_2 blow dry.
Metallise	- Au 10 nm, Zn 5 nm, Au 150 nm.
Lift-off	- 2 hrs 40 °C acetone, 5 min IPA, N_2 blow dry.
RTA	- N_2 ambient, 400 °C, 3 s.

A.1.2 Isolation

Clean Substrate	- 5 min 40 °C acetone, 5 min IPA, N_2 blow dry.
Spin Resist	- Microposit S1818, 4 krpm, 30 s.
Bake	- 90 °C oven, 30 min.
MA6 Exposure	- 3.8 s, hard contact.
Develop	- 75 s Microposit Developer:H_2O (1:1), 30 s H_2O rinse, N_2 blow dry.

Dehydration Bake	- 200 °C hotplate, 5 min.
Spin Resist	- SU-8, 500 rpm, 30 s $\longrightarrow$ 3 krpm, 30 s.
Bake	- 65 °C hotplate, 1 min. 95 °C hotplate, 1 min.
MA6 Exposure	- 4.5 s, hard contact.
Develop	- 1 min Microposit EC Solvent, 30 s IPA rinse, N_2 blow dry.
Bake	- 150 °C hotplate, 1 min.
De-oxidise	- 30 s HCl:H_2O (1:4), 30 s H_2O rinse.
Etch	- 80 s H_3PO_4:H_2O_2:H_2O (1:1:10), 30 s H_2O rinse, N_2 blow dry.
Remove Resist	- >24 hrs Microposit Remover 1165, IPA rinse, N_2 blow dry.

A.2 Baseline InGaAs-Channel Devices (EBL)

A.2.1 Marker/Gate

Clean Substrate	- 5 min 40 °C acetone, 5 min IPA, N_2 blow dry.
Spin Resist	- 8% 2010 PMMA, 5 krpm, 60 s.
Bake	- 180 °C oven, 1 hr.
Spin Resist	- 4% 2041 PMMA, 5 krpm, 60 s.
Bake	- 180 °C oven, 1 hr.
VB6 Exposure	- Global markers: Dose 330 μC/cm^2, 64 nA beam, 33 nm spot, VRU 40.
	- Penrose markers: Dose 360 μC/cm^2, 1 nA beam, 4 nm spot, VRU 4.
	- L_g=100 nm: Dose 1150 μC/cm^2, 4 nA beam, 9 nm spot, VRU 20.
	- L_g=0.3−2 μm: Dose 725 μC/cm^2, 4 nA beam, 9 nm spot, VRU 20.
	- L_g=20 μm: Dose 330 μC/cm^2, 64 nA beam, 33 nm spot, VRU 40.
Develop	- 60 s IPA:MIBK (2.5:1) at 23 °C, 30 s IPA rinse, N_2 blow dry.
Ash	- O_2 plasma, 40 W, 60 s.
Metallise	- Al 20 nm, Au 100 nm.
Lift-off	- 2 hrs 40 °C acetone, 5 min IPA, N_2 blow dry.

A.2.2 Ohmic

Clean Substrate	- 5 min 40 °C acetone, 5 min IPA, N_2 blow dry.
Spin Resist	- 12% 2010 PMMA, 5 krpm, 60 s.
Bake	- 180 °C oven, 1 hr.
Spin Resist	- 4% 2041 PMMA, 5 krpm, 60 s.
Bake	- 180 °C oven, 1 hr.
VB6 Exposure	- Dose 305 μC/cm^2, 64 nA beam, 33 nm spot, VRU 40.
Develop	- 60 s IPA:MIBK (2:1) at 23 °C, 30 s IPA rinse, N_2 blow dry.
Ash	- O_2 plasma, 40 W, 60 s.
Al_2O_3 Etch	- 2 min 45 s AZ 400k Developer, 30 s H_2O rinse, N_2 blow dry.
Metallise	- Au 5 nm, Zn 40 nm, Au 150 nm.
Lift-off	- 2 hrs 40 °C acetone, 5 min IPA, N_2 blow dry.
Furnace Anneal	- FG (5% H_2: 95% N_2) ambient, 380 °C, 2 min.

A.3 Scaled Lside InGaAs-Channel Devices (EBL)

A.3.1 Marker/Gate

Clean Substrate	- 5 min 40 °C acetone, 5 min IPA, N_2 blow dry.
Spin Resist	- 8% 2010 PMMA, 5 krpm, 60 s.
Bake	- 180 °C oven, 1 hr.
Spin Resist	- 4% 2041 PMMA, 5 krpm, 60 s.
Bake	- 180 °C oven, 1 hr.
VB6 Exposure	- Global markers: Dose 330 μC/cm^2, 64 nA beam, 33 nm spot, VRU 40.
	- Penrose markers: Dose 360 μC/cm^2, 1 nA beam, 4 nm spot, VRU 4.
	- L_g=100 nm: Dose 1150 μC/cm^2, 4 nA beam, 9 nm spot, VRU 20.
	- L_g=0.3−2 μm: Dose 725 μC/cm^2, 4 nA beam, 9 nm spot, VRU 20.
	- L_g=20 μm: Dose 330 μC/cm^2, 64 nA beam, 33 nm spot, VRU 40.
Develop	- 60 s IPA:MIBK (2.5:1) at 23 °C, 30 s IPA rinse, N_2 blow dry.
Ash	- O_2 plasma, 40 W, 60 s.
Metallise	- Al 20 nm, Au 100 nm.
Lift-off	- 2 hrs 40 °C acetone, 5 min IPA, N_2 blow dry.

A.3.2 Stripe Ohmic (2 μm×100 μm)

Clean Substrate	- 5 min 40 °C acetone, 5 min IPA, N_2 blow dry.
Spin Resist	- 4% 2010 PMMA, 3 krpm, 60 s.
Bake	- 180 °C oven, 1 hr.
Spin Resist	- 2.5% 2041 PMMA, 5 krpm, 60 s.
Bake	- 180 °C oven, 1 hr.
VB6 Exposure	- L_{side}=70 nm: Dose 200−290 μC/cm^2, 8 nA beam, 12 nm spot, VRU 12.
	- L_{side}=200 nm: Dose 280 − 330 μC/cm^2, 8 nA beam, 12 nm spot, VRU 12.
	- L_{side}=465 nm: Dose 230−330 μC/cm^2, 64 nA beam, 33 nm spot, VRU 12.
	- L_{side}=1 μm: Dose 305 μC/cm^2, 8 nA beam, 12 nm spot, VRU 12.

Develop	- 60 s IPA:MIBK (2:1) at 23 °C, 30 s IPA rinse, N_2 blow dry.
Ash	- O_2 plasma, 40 W, 60 s.
Al_2O_3 Etch	- 2 min 45 s AZ 400k Developer, 30 s H_2O rinse, N_2 blow dry.
Metallise	- Au 5 nm, Zn 40 nm, Au 100 nm.
Lift-off	- 2 hrs 40 °C acetone, 5 min IPA, N_2 blow dry.
Furnace Anneal	- FG (5% H_2: 95% N_2) ambient, 380 °C, 2 min.

A.3.3 Large Ohmic (150 μm×100 μm)

Clean Substrate	- 5 min 40 °C acetone, 5 min IPA, N_2 blow dry.
Spin Resist	- 12% 2010 PMMA, 5 krpm, 60 s.
Bake	- 180 °C oven, 1 hr.
Spin Resist	- 4% 2041 PMMA, 5 krpm, 60 s.
Bake	- 180 °C oven, 1 hr.
VB6 Exposure	- Dose 305 μC/cm^2, 64 nA beam, 33 nm spot, VRU 40.
Develop	- 60 s IPA:MIBK (2:1) at 23 °C, 30 s IPA rinse, N_2 blow dry.
Ash	- O_2 plasma, 40 W, 60 s.
Metallise	- Ti 30 nm, Pt 50 nm, Au 100 nm.
Lift-off	- 2 hrs 40 °C acetone, 5 min IPA, N_2 blow dry.

A.4 Baseline InGaSb-Channel Devices (EBL)

A.4.1 Marker/Gate

Clean Substrate	- 5 min 40 °C acetone, 5 min IPA, N_2 blow dry.
Spin Resist	- 8% 2010 PMMA, 5 krpm, 60 s.
Bake	- 180 °C oven, 1 hr.
Spin Resist	- 4% 2041 PMMA, 5 krpm, 60 s.
Bake	- 180 °C oven, 1 hr.
VB6 Exposure	- Global markers: Dose 330 μC/cm^2, 64 nA beam, 33 nm spot, VRU 40.
	- Penrose markers: Dose 360 μC/cm^2, 1 nA beam, 4 nm spot, VRU 4.
	- L_g=100 nm: Dose 1150 μC/cm^2, 4 nA beam, 9 nm spot, VRU 20.
	- L_g=0.3−2 μm: Dose 725 μC/cm^2, 4 nA beam, 9 nm spot, VRU 20.
	- L_g=20 μm: Dose 330 μC/cm^2, 64 nA beam, 33 nm spot, VRU 40.
Develop	- 60 s IPA:MIBK (2.5:1) at 23 °C, 30 s IPA rinse, N_2 blow dry.
Ash	- O_2 plasma, 40 W, 60 s.
Metallise	- Al 20 nm, Au 100 nm.
Lift-off	- 2 hrs 40 °C acetone, 5 min IPA, N_2 blow dry.

A.4.2 Ohmic

Clean Substrate	- 5 min 40 °C acetone, 5 min IPA, N_2 blow dry.
Spin Resist	- 12% 2010 PMMA, 5 krpm, 60 s.
Bake	- 180 °C oven, 1 hr.
Spin Resist	- 4% 2041 PMMA, 5 krpm, 60 s.
Bake	- 180 °C oven, 1 hr.
VB6 Exposure	- Dose 305 μC/cm^2, 64 nA beam, 33 nm spot, VRU 40.
Develop	- 60 s IPA:MIBK (2:1) at 23 °C, 30 s IPA rinse, N_2 blow dry.
Ash	- O_2 plasma, 40 W, 60 s.
Al_2O_3 Etch	- 5 min AZ 400k Developer, 30 s H_2O rinse, N_2 blow dry.
Metallise	- Au 5 nm, Zn 40 nm, Au 100 nm.
Lift-off	- 2 hrs 40 °C acetone, 5 min IPA, N_2 blow dry.
Furnace Anneal	- FG (5% H_2: 95% N_2) ambient, 340 °C, 2 min.

Bibliography

[1] D. Kahng and A. M.M., "Silicon-silicon dioxide field induced surface device," in *IRE-AIEE Solid-State Device Research Conference*, 1960.

[2] G. Moore, "Progress in digital integrated electronics," in *IEEE International Electron Devices Meeting (IEDM) Technical Digest*, 1975, pp. 11–13.

[3] R. H. Dennard, F. H. Gaensslen, V. L. Rideout, E. Bassous, and A. R. LeBlanc, "Design of ion-implanted MOSFET's with very small physical dimensions," *IEEE Journal of Solid-State Circuits*, vol. 9, no. 5, pp. 256–268, 1974.

[4] H. Iwai, "Roadmap for 22 nm and beyond (invited paper)," *Microelectronic Engineering*, vol. 86, no. 79, pp. 1520–1528, 2009.

[5] E. J. Nowak, "Maintaining the benefits of CMOS scaling when scaling bogs down," *IBM Journal of Research and Development*, vol. 46, no. 2.3, pp. 169–180, 2002.

[6] D. J. Frank, "Power-constrained CMOS scaling limits," *IBM Journal of Research and Development*, vol. 46, no. 2.3, pp. 235–244, 2002.

[7] K. C. Saraswat, D. Kim, T. Krishnamohan, and A. Pethe, "Performance limitations of Si bulk CMOS and alternatives for future ULSI," *Journal of the Indian Institute of Science*, vol. 87, no. 3, pp. 387–400, 2007.

[8] Y.-B. Kim, "Challenges for nanoscale MOSFETs and emerging nanoelectronics," *Transactions on Electrical and Electronic Materials*, vol. 11, no. 3, pp. 93–105, 2010.

[9] S. E. Thompson, M. Armstrong, C. Auth, M. Alavi, M. Buehler, R. Chau, S. Cea, T. Ghani, G. Glass, T. Hoffman, C. H. Jan, C. Kenyon, J. Klaus, K. Kuhn, Z. Ma, B. Mcintyre, K. Mistry, A. Murthy, B. Obradovic, R. Nagisetty, P. Nguyen, S. Sivakumar, R. Shaheed, L. Shifren, B. Tufts, S. Tyagi, M. Bohr, and Y. El-Mansy, "A 90-nm logic technology featuring strained-silicon," *IEEE Transactions on Electron Devices*, vol. 51, no. 11, pp. 1790–1797, 2004.

[10] K. Mistry, C. Allen, C. Auth, B. Beattie, D. Bergstrom, M. Bost, M. Brazier, M. Buehler, A. Cappellani, R. Chau, C.-H. Choi, G. Ding, K. Fischer, T. Ghani, R. Grover, W. Han, D. Hanken, M. Hattendorf, J. He, J. Hicks, R. Huessner, D. Ingerly, P. Jain, R. James, L. Jong, S. Joshi, C. Kenyon, K. Kuhn, K. Lee, H. Liu, J. Maiz, B. Mclntyre, P. Moon, J. Neirynck, S. Pae, C. Parker, D. Parsons, C. Prasad, L. Pipes, M. Prince, P. Ranade, T. Reynolds, J. Sandford, L. Shifren, J. Sebastian, J. Seiple, D. Simon, S. Sivakumar, P. Smith, C. Thomas, T. Troeger, P. Vandervoorn, S. Williams, and K. Zawadzki, "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 (IEDM) Technical Digest*, 2007, pp. 247–250.

[11] Intel, "Intel 22nm 3-D tri-gate transistor technology," 2016. [Online]. Available: http://newsroom.intel.co.uk/press-kits/intel-22nm-3-d-tri-gate-transistor-technology/

[12] S. Natarajan, M. Agostinelli, S. Akbar, M. Bost, A. Bowonder, V. Chikarmane, S. Chouksey, A. Dasgupta, K. Fischer, Q. Fu, T. Ghani, M. Giles, S. Govindaraju, R. Grover, W. Han, D. Hanken, E. Haralson, M. Haran, M. Heckscher, R. Heussner, P. Jain, R. James, R. Jhaveri, I. Jin, H. Kam, E. Karl, C. Kenyon, M. Liu, Y. Luo, R. Mehandru, S. Morarka, L. Neiberg, P. Packan, A. Paliwal, C. Parker, P. Patel, R. Patel, C. Pelto, L. Pipes, P. Plekhanov, M. Prince, S. Rajamani, J. Sandford, B. Sell, S. Sivakumar, P. Smith, B. Song, K. Tone, T. Troeger, J. Wiedemer, M. Yang, and K. Zhang, "A 14nm logic technology featuring 2nd-generation FinFET, air-gapped interconnects, self-aligned double patterning and a 0.0588 μm^2 SRAM cell size," in *Electron Devices Meeting (IEDM), 2014 IEEE International*, 2014, pp. 71–73.

[13] H. Riel, L.-E. Wernersson, M. Hong, and J. A. del Alamo, "IIIV compound semiconductor transistors from planar to nanowire structures," *MRS Bulletin*, vol. 39, pp. 668–677, 8 2014.

[14] J. del Alamo, "Nanometre-scale electronics with IIIV compound semiconductors," *Nature*, vol. 479, no. 7373, pp. 317–323, 2011.

[15] S. Oktyabrsky, "Materials and Technologies for III-V MOSFETs," in *Fundamentals of III-V Semiconductor MOSFETs*, S. Oktyabrsky and P. Ye, Eds. Springer US, 2010, pp. 195–250.

[16] C. Hinkle, E. Vogel, P. Ye, and R. Wallace, "Interfacial chemistry of oxides on $In_xGa_{(1-x)}As$ and implications for MOSFET applications," *Current Opinion in Solid State and Materials Science*, vol. 15, no. 5, pp. 188–207, 2011.

[17] W. E. Spicer, I. Lindau, P. Skeath, and C. Y. Su, "Unified defect model and beyond," *Journal of Vacuum Science & Technology*, vol. 17, no. 5, pp. 1019–1027, 1980.

[18] M. Passlack, M. Hong, J. P. Mannaerts, J. R. Kwo, and L. W. Tu, "Recombination velocity at oxide-GaAs interfaces fabricated by in situ molecular beam epitaxy," *Applied Physics Letters*, vol. 68, no. 25, pp. 3605–3607, 1996.

[19] M. Passlack, M. Hong, and J. P. Mannaerts, "Quasistatic and high frequency capacitance-voltage characterization of Ga_2O_3-GaAs structures fabricated by in situ molecular beam epitaxy," *Applied Physics Letters*, vol. 68, no. 8, pp. 1099–1101, 1996.

[20] M. Passlack, M. Hong, J. P. Mannaerts, S. N. G. Chu, R. L. Opila, and N. Moriya, "In-situ Ga_2O_3 process for GaAs inversion/accumulation device and surface passivation applications," in *IEEE International Electron Devices Meeting (IEDM) Technical Digest*, 1995, pp. 383–386.

[21] P. D. Ye, G. D. Wilk, B. Yang, J. Kwo, S. N. G. Chu, S. Nakahara, H.-J. L. Gossmann, J. P. Mannaerts, M. Hong, K. K. Ng, and J. Bude, "GaAs metal-oxide-semiconductor field-effect transistor with nanometer-thin dielectric grown by atomic layer deposition," *Applied Physics Letters*, vol. 83, no. 1, pp. 180–182, 2003.

[22] P. D. Ye, G. D. Wilk, J. Kwo, B. Yang, H. J. L. Gossmann, M. Frei, S. N. G. Chu, J. P. Mannaerts, M. Sergent, M. Hong, K. K. Ng, and J. Bude, "GaAs MOSFET with oxide gate dielectric grown by atomic layer deposition," *IEEE Electron Device Letters*, vol. 24, no. 4, pp. 209–211, 2003.

[23] R. Suzuki, N. Taoka, M. Yokoyama, S. Lee, S. H. Kim, T. Hoshii, T. Yasuda, W. Jevasuwan, T. Maeda, O. Ichikawa, N. Fukuhara, M. Hata, M. Takenaka, and S. Takagi, "1-nm-capacitance-equivalent-thickness $HfO_2/Al_2O_3/InGaAs$ metal-oxide-semiconductor structure with low interface trap density and low gate leakage current density," *Applied Physics Letters*, vol. 100, no. 13, p. 132906, 2012.

[24] A. D. Carter, W. J. Mitchell, B. J. Thibeault, J. J. M. Law, and M. J. W. Rodwell, "Al_2O_3 growth on (100) $In_{0.53}Ga_{0.47}As$ initiated by cyclic trimethylaluminum and hydrogen plasma exposures," *Applied Physics Express*, vol. 4, no. 9, p. 091102, 2011.

[25] V. Chobpattana, J. Son, J. J. M. Law, R. Engel-Herbert, C.-Y. Huang, and S. Stemmer, "Nitrogen-passivated dielectric/InGaAs interfaces with sub-nm equivalent oxide thickness and low interface trap densities," *Applied Physics Letters*, vol. 102, no. 2, p. 022907, 2013.

[26] E. O'Connor, B. Brennan, V. Djara, K. Cherkaoui, S. Monaghan, S. B. Newcomb, R. Contreras, M. Milojevic, G. Hughes, M. E. Pemble, R. M. Wallace, and P. K. Hurley, "A systematic study of $(NH_4)_2S$ passivation (22%, 10%, 5%, or 1%) on the interface properties of the $Al_2O_3/In_{0.53}Ga_{0.47}As/InP$ system for n-type and p-type $In_{0.53}Ga_{0.47}As$ epitaxial layers," *Journal of Applied Physics*, vol. 109, no. 2, p. 024101, 2011.

[27] E. O'Connor, R. D. Long, K. Cherkaoui, K. K. Thomas, F. Chalvet, I. M. Povey, M. E. Pemble, P. K. Hurley, B. Brennan, G. Hughes, and S. B. Newcomb, "In situ H_2S passivation of $In_{0.53}Ga_{0.47}As/InP$ metal-oxide-semiconductor capacitors with atomic-layer deposited HfO_2 gate dielectric," *Applied Physics Letters*, vol. 92, no. 2, p. 022902, 2008.

[28] A. Alian, G. Brammertz, C. Merckling, A. Firrincieli, W.-E. Wang, H. C. Lin, M. Caymax, M. Meuris, K. D. Meyer, and M. Heyns, "Ammonium sulfide vapor passivation of $In_{0.53}Ga_{0.47}As$ and InP surfaces," *Applied Physics Letters*, vol. 99, no. 11, p. 112114, 2011.

[29] E. J. Kim, E. Chagarov, J. Cagnon, Y. Yuan, A. C. Kummel, P. M. Asbeck, S. Stemmer, K. C. Saraswat, and P. C. McIntyre, "Atomically abrupt and unpinned $Al_2O_3/In_{0.53}Ga_{0.47}As$ interfaces: experiment and simulation," *Journal of Applied Physics*, vol. 106, no. 12, p. 124508, 2009.

[30] J. Ahn, I. Geppert, M. Gunji, M. Holland, I. Thayne, M. Eizenberg, and P. C. McIntyre, "Titania/alumina bilayer gate insulators for InGaAs metal-oxide-semiconductor devices," *Applied Physics Letters*, vol. 99, no. 23, p. 232902, 2011.

[31] I. Ok, H. Kim, M. Zhang, T. Lee, F. Zhu, L. Yu, S. Koveshnikov, W. Tsai, V. Tokranov, M. Yakimov, S. Oktyabrsky, and J. C. Lee, "Self-aligned n- and p-channel GaAs MOSFETs on undoped and p-type substrates using HfO_2 and silicon interface passivation layer," in *IEEE International Electron Devices Meeting (IEDM) Technical Digest*, 2006, pp. 1–4.

[32] M. Radosavljevic, B. Chu-Kung, S. Corcoran, G. Dewey, M. Hudait, J. Fastenau, J. Kavalieros, W. K. Liu, D. Lubyshev, M. Metz, K. Millard, N. Mukherjee, W. Rachmady, U. Shah, and R. Chau, "Advanced high-k gate dielectric for high-performance short-channel

$In_{0.7}Ga_{0.3}As$ quantum well field effect transistors on silicon substrate for low power logic applications," in *IEEE International Electron Devices Meeting (IEDM) Technical Digest*, 2009, pp. 319–322.

[33] D. Shahrjerdi, T. Rotter, G. Balakrishnan, D. Huffaker, E. Tutuc, and S. Banerjee, "Fabrication of self-aligned enhancement-mode $In_{0.53}Ga_{0.47}As$ MOSFETs with $TaN/HfO_2/AlN$ gate stack," *Electron Device Letters, IEEE*, vol. 29, no. 6, pp. 557–560, 2008.

[34] C.-W. Cheng, G. Apostolopoulos, and E. A. Fitzgerald, "The effect of interface processing on the distribution of interfacial defect states and the C-V characteristics of III-V metal-oxide-semiconductor field effect transistors," *Journal of Applied Physics*, vol. 109, no. 2, p. 023714, 2011.

[35] H. Zhao, J. H. Yum, Y.-T. Chen, and J. C. Lee, "$In_{0.53}Ga_{0.47}As$ n-metal-oxide-semiconductor field effect transistors with atomic layer deposited Al_2O_3, HfO_2, and $LaAlO_3$ gate dielectrics," *Journal of Vacuum Science Technology B: Microelectronics and Nanometer Structures*, vol. 27, no. 4, pp. 2024–2027, 2009.

[36] P. K. Hurley, E. O'Connor, S. Monaghan, R. Long, A. O'Mahony, I. M. Povey, K. Cherkaoui, J. MacHale, A. Quinn, G. Brammertz, M. M. Heyns, S. Newcomb, V. V. Afanas'ev, A. Sonnet, R. Galatage, N. Jivani, E. Vogel, R. M. Wallace, and M. Pemble, "Structural and electrical properties of HfO_2/n-$In_xGa_{1-x}As$ structures (x: 0, 0.15, 0.3 and 0.53)," *ECS Transactions*, vol. 25, no. 6, pp. 113–127, 2009.

[37] N. Goel, P. Majhi, C. O. Chui, W. Tsai, D. Choi, and J. S. Harris, "InGaAs metal-oxide-semiconductor capacitors with HfO_2 gate dielectric grown by atomic-layer deposition," *Applied Physics Letters*, vol. 89, no. 16, p. 163517, 2006.

[38] H. Zhao, J. Huang, Y.-T. Chen, J. H. Yum, Y. Wang, F. Zhou, F. Xue, and J. C. Lee, "Effects of gate-first and gate-last process on interface quality of $In_{0.53}Ga_{0.47}As$ metal-oxide-semiconductor capacitors using atomic-layer-deposited Al_2O_3 and HfO_2 oxides," *Applied Physics Letters*, vol. 95, no. 25, p. 253501, 2009.

[39] R. Engel-Herbert, Y. Hwang, J. Cagnon, and S. Stemmer, "Metal-oxide-semiconductor capacitors with ZrO_2 dielectrics grown on $In_{0.53}Ga_{0.47}As$ by chemical beam deposition," *Applied Physics Letters*, vol. 95, no. 6, p. 062908, 2009.

[40] M. Barth, G. Bruce Rayner, S. McDonnell, R. M. Wallace, B. R. Bennett, R. Engel-Herbert, and S. Datta, "High quality HfO2/p-GaSb(001) metal-oxide-semiconductor capacitors with 0.8 nm equivalent oxide thickness," *Applied Physics Letters*, vol. 105, no. 22, p. 222103, 2014.

[41] S. Koveshnikov, N. Goel, P. Majhi, H. Wen, M. B. Santos, S. Oktyabrsky, V. Tokranov, R. Kambhampati, R. Moore, F. Zhu, J. Lee, and W. Tsai, "$In_{0.53}Ga_{0.47}As$ based metal oxide semiconductor capacitors with atomic layer deposition ZrO_2 gate oxide demonstrating low gate leakage current and equivalent oxide thickness less than 1nm," *Applied Physics Letters*, vol. 92, no. 22, p. 222904, 2008.

[42] S. W. Chang, X. Li, R. Oxland, S. W. Wang, C. H. Wang, R. Contreras-Guerrero, K. K. Bhuwalka, G. Doornbos, T. Vasen, M. C. Holland, G. Vellianitis, M. J. H. van Dal, B. Duriez, M. Edirisooriya, J. S. Rojas-Ramirez, P. Ramvall, S. Thoms, U. Peralagu, C. H. Hsieh, Y. S.

Chang, K. M. Yin, E. Lind, L. E. Wernersson, R. Droopad, I. Thayne, M. Passlack, and C. H. Diaz, "InAs N-MOSFETs with record performance of Ion = 600 μA/μm at Ioff = 100 nA/μm (Vd = 0.5 V)," in *IEEE International Electron Devices Meeting (IEDM) Technical Digest*, Dec 2013, pp. 417–420.

[43] J. Lin, X. Zhao, T. Yu, D. A. Antoniadis, and J. A. D. Alamo, "A new self-aligned quantum-well MOSFET architecture fabricated by a scalable tight-pitch process," in *IEEE International Electron Devices Meeting (IEDM) Technical Digest*, 2013, pp. 421–424.

[44] J. Lin, D. A. Antoniadis, and J. A. del Alamo, "Novel intrinsic and extrinsic engineering for high-performance high-density self-aligned ingaas mosfets: Precise channel thickness control and sub-40-nm metal contacts," in *Electron Devices Meeting (IEDM), 2014 IEEE International*, Dec 2014, pp. 574–577.

[45] S. Lee, C. Y. Huang, A. D. Carter, D. C. Elias, J. J. M. Law, V. Chobpattana, S. Krmer, B. J. Thibeault, W. Mitchell, S. Stemmer, A. C. Gossard, and M. J. W. Rodwell, "Record extrinsic transconductance (2.45 mS/μm at VDS = 0.5 V) InAs/$In_{0.53}Ga_{0.47}As$ channel MOSFETs using MOCVD source-drain regrowth," in *Symposium on VLSI Technology Digest Technical Papers*, 2013, pp. T246–T247.

[46] S. Lee, V. Chobpattana, C. Y. Huang, B. J. Thibeault, W. Mitchell, S. Stemmer, A. C. Gossard, and M. J. W. Rodwell, "Record I_{on} (0.50 mA/μm at V_{DD}=0.5 V and I_{off}=100 nA/μm) 25 nm-gate-length ZrO_2/InAs/InAlAs MOSFETs," in *Symposium on VLSI Technology Digest Technical Papers*, 2014, pp. 1–2.

[47] I. Thayne, S. Bentley, M. Holland, W. Jansen, X. Li, D. Macintyre, S. Thoms, B. Shin, J. Ahn, and P. McIntyre, "III-V nMOSFETs - some issues associated with roadmap worthiness (invited)," *Microelectronic Engineering*, vol. 88, no. 7, pp. 1070–1075, 2011.

[48] R. J. Baker, "CMOS fabrication," in *CMOS Circuit Design, Layout, and Simulation*, 3rd ed. Wiley-IEEE Press, 2010, pp. 161–212.

[49] R. J. Hill, J. Oh, C. Park, J. Barnett, J. Price, J. Huang, N. Goel, W. Y. Loh, P. Kirsch, P. Majhi, and R. Jammy, "CMOS Scaling with III-V Channels for Improved Performance and Low Power," *ECS Transactions*, vol. 35, no. 3, pp. 335–344, 2011.

[50] V. Deshpande, V. Djara, E. O'Connor, P. Hashemi, K. Balakrishnan, M. Sousa, D. Caimi, A. Olziersky, L. Czornomaz, and J. Fompeyrine, "Advanced 3D monolithic hybrid CMOS with sub-50 nm gate inverters featuring replacement metal gate (RMG)-InGaAs nFETs on SiGe-OI Fin pFETs," in *IEEE International Electron Devices Meeting (IEDM) Technical Digest*, 2015, pp. 209–212.

[51] R. Oxland, S. W. Chang, X. Li, S. W. Wang, G. Radhakrishnan, W. Priyantha, M. J. H. Van Dal, C. Hsieh, G. Vellianitis, G. Doornbos, K. Bhuwalka, B. Duriez, I. Thayne, R. Droopad, M. Passlack, C. Diaz, and Y. C. Sun, "An ultralow-resistance ultrashallow metallic source/drain contact scheme for III-V NMOS," *Electron Device Letters, IEEE*, vol. 33, no. 4, pp. 501–503, 2012.

[52] M. Abraham, S.-Y. Yu, W. H. Choi, R. T. P. Lee, and S. E. Mohney, "Very low resistance alloyed Ni-based ohmic contacts to InP-capped and uncapped n^+-$In_{0.53}Ga_{0.47}As$," *Journal of Applied Physics*, vol. 116, no. 16, p. 164506, 2014.

[53] S. Kim, M. Yokoyama, R. Nakane, O. Ichikawa, T. Osada, M. Hata, M. Takenaka, and S. Takagi, "High-performance InAs-on-insulator n-MOSFETs with Ni-InGaAs S/D realized by contact resistance reduction technology," *IEEE Transactions on Electron Devices*, vol. 60, no. 10, pp. 3342–3350, 2013.

[54] S. Kim, M. Yokoyama, N. Taoka, R. Nakane, T. Yasuda, O. Ichikawa, N. Fukuhara, M. Hata, M. Takenaka, and S. Takagi, "$In_{0.53}Ga_{0.47}As$ metal-oxide-semiconductor field-effect transistors with self-aligned metal source/drain using Co-InGaAs alloys," *Applied Physics Letters*, vol. 100, no. 7, p. 073504, 2012.

[55] J. D. Yearsley, J. C. Lin, E. Hwang, S. Datta, and S. E. Mohney, "Ultra low-resistance palladium silicide ohmic contacts to lightly doped n-InGaAs," *Journal of Applied Physics*, vol. 112, no. 5, p. 054510, 2012.

[56] W. Lu, A. Guo, A. Vardi, and J. A. del Alamo, "A test structure to characterize nanoscale ohmic contacts in III-V MOSFETs," *IEEE Electron Device Letters*, vol. 35, no. 2, pp. 178–180, 2014.

[57] A. Baraskar, M. A. Wistey, V. Jain, E. Lobisser, U. Singisetti, G. Burek, Y. J. Lee, B. Thibeault, A. Gossard, and M. Rodwell, "Ex situ ohmic contacts to n-InGaAs," *Journal of Vacuum Science & Technology B*, vol. 28, no. 4, pp. C5I7–C5I9, 2010.

[58] A. K. Baraskar, M. A. Wistey, V. Jain, U. Singisetti, G. Burek, B. J. Thibeault, Y. J. Lee, A. C. Gossard, and M. J. W. Rodwell, "Ultralow resistance, nonalloyed ohmic contacts to n-InGaAs," *Journal of Vacuum Science & Technology B*, vol. 27, no. 4, pp. 2036–2039, 2009.

[59] M. K. Hudait, G. Dewey, S. Datta, J. M. Fastenau, J. Kavalieros, W. K. Liu, D. Lubyshev, R. Pillarisetty, W. Rachmady, M. Radosavljevic, T. Rakshit, and R. Chau, "Heterogeneous integration of enhancement mode $In_{0.7}Ga_{0.3}As$ quantum well transistor on silicon substrate using thin ($\leq 2\,\mu$m) composite buffer architecture for high-speed and low-voltage (0.5V) logic applications," in *IEEE International Electron Devices Meeting (IEDM) Technical Digest*, 2007, pp. 625–628.

[60] C. W. Tang, Z. Zhong, and K. M. Lau, "Hetero-epitaxy of III-V compounds by MOCVD on silicon substrates," *ECS Transactions*, vol. 28, no. 5, pp. 227–231, 2010.

[61] Y. Q. Wu, M. Xu, P. D. Ye, Z. Cheng, J. Li, J.-S. Park, J. Hydrick, J. Bai, M. Carroll, J. G. Fiorenza, and A. Lochtefeld, "Atomic-layer-deposited Al_2O_3/GaAs metal-oxide-semiconductor field-effect transistor on Si substrate using aspect ratio trapping technique," *Applied Physics Letters*, vol. 93, no. 24, p. 242106, 2008.

[62] C. Merckling, N. Waldron, S. Jiang, W. Guo, O. Richard, B. Douhard, A. Moussa, D. Vanhaeren, H. Bender, N. Collaert, M. Heyns, A. Thean, M. Caymax, and W. Vandervorst, "Selective area growth of InP in shallow trench isolation on large scale Si(001) wafer using defect confinement technique," *Journal of Applied Physics*, vol. 114, no. 3, p. 033708, 2013.

[63] N. Waldron, G. Wang, N. D. Nguyen, T. Orzali, C. Merckling, G. Brammertz, P. Ong, G. Winderickx, G. Hellings, G. Eneman, M. Caymax, M. Meuris, N. Horiguchi, and A. Thean, "Integration of InGaAs channel n-MOS devices on 200mm Si wafers using the aspect-ratio-trapping technique," *ECS Transactions*, vol. 45, no. 4, pp. 115–128, 2012.

[64] M. Yokoyama, R. Iida, S. H. Kim, N. Taoka, Y. Urabe, T. Yasuda, H. Takagi, H. Yamada, N. Fukuhara, M. Hata, M. Sugiyama, Y. Nakano, M. Takenaka, and S. Takagi, "Extremely-thin-body InGaAs-on-insulator MOSFETs on Si fabricated by direct wafer bonding," in *IEEE International Electron Devices Meeting (IEDM) Technical Digest*, 2010, pp. 46–49.

[65] L. Czornomaz, N. Daix, D. Caimi, M. Sousa, R. Erni, M. D. Rossell, M. El-Kazzi, C. Rossel, C. Marchiori, E. Uccelli, M. Richter, H. Siegwart, and J. Fompeyrine, "An integration path for gate-first UTB III-V-on-insulator MOSFETs with silicon, using direct wafer bonding and donor wafer recycling," in *IEEE International Electron Devices Meeting (IEDM) Technical Digest*, 2012, pp. 517–520.

[66] H. Fang, M. Madsen, C. Carraro, K. Takei, H. S. Kim, E. Plis, S.-Y. Chen, S. Krishna, Y.-L. Chueh, R. Maboudian, and A. Javey, "Strain engineering of epitaxially transferred, ultrathin layers of III-V semiconductor on insulator," *Applied Physics Letters*, vol. 98, no. 1, p. 012111, 2011.

[67] K. Takei, M. Madsen, H. Fang, R. Kapadia, S. Chuang, H. S. Kim, C.-H. Liu, E. Plis, J. Nah, S. Krishna, Y.-L. Chueh, J. Guo, and A. Javey, "Nanoscale InGaSb heterostructure membranes on Si substrates for high hole mobility transistors," *Nano Letters*, vol. 12, no. 4, pp. 2060–2066, 2012.

[68] Ioffe, "Physical properties of semiconductors," 1998. [Online]. Available: http://www.ioffe.ru/SVA/NSM/Semicond/index.html

[69] R. Pillarisetty, B. Chu-Kung, S. Corcoran, G. Dewey, J. Kavalieros, H. Kennel, R. Kotlyar, V. Le, D. Lionberger, M. Metz, N. Mukherjee, J. Nah, W. Rachmady, M. Radosavljevic, U. Shah, S. Taft, H. Then, N. Zelick, and R. Chau, "High mobility strained germanium quantum well field effect transistor as the p-channel device option for low power (Vcc=0.5V) III-V CMOS architecture," in *IEEE International Electron Devices Meeting (IEDM) Technical Digest*, 2010, pp. 150–153.

[70] M. R. Caymax, F. Leys, J. Mitard, K. Martens, L. Yang, G. Pourtois, W. Vandervorst, M. Meuris, and R. Loo, "The influence of the epitaxial growth process parameters on layer characteristics and device performance in Si-passivated Ge pMOSFETs," *ECS Transactions*, vol. 19, no. 1, pp. 183–194, 2009.

[71] L. Hutin, C. L. Royer, J. F. Damlencourt, J. M. Hartmann, H. Grampeix, V. Mazzocchi, C. Tabone, B. Previtali, A. Pouydebasque, M. Vinet, and O. Faynot, "GeOI pMOSFETs scaled down to 30-nm gate length with record off-state current," *IEEE Electron Device Letters*, vol. 31, no. 3, pp. 234–236, 2010.

[72] J. Mitard, B. D. Jaeger, F. E. Leys, G. Hellings, K. Martens, G. Eneman, D. P. Brunco, R. Loo, J. C. Lin, D. Shamiryan, T. Vandeweyer, G. Winderickx, E. Vrancken, C. H. Yu, K. D. Meyer, M. Caymax, L. Pantisano, M. Meuris, and M. M. Heyns, "Record I_{ON}/I_{OFF} performance for 65nm Ge pMOSFET and novel Si passivation scheme for improved EOT scalability," in *IEEE International Electron Devices Meeting (IEDM) Technical Digest*, 2008, pp. 1–4.

[73] P. Nagaiah, V. Tokranov, M. Yakimov, and S. Oktyabrsky, "Strained quantum wells for p-channel InGaAs CMOS," in *Symposium A Performance and Reliability of Semiconductor Devices*, ser. MRS Proceedings, vol. 1108, 2008.

[74] B. R. Bennett, T. F. Chick, M. G. Ancona, and J. B. Boos, "Enhanced hole mobility and density in GaSb quantum wells," *Solid-State Electronics*, vol. 79, pp. 274–280, 2013.

[75] B. R. Bennett, M. G. Ancona, J. B. Boos, C. B. Canedy, and S. A. Khan, "Strained GaSb/AlAsSb quantum wells for p-channel field-effect transistors," *Journal of Crystal Growth*, vol. 311, no. 1, pp. 47–53, 2008.

[76] B. R. Bennett, M. G. Ancona, J. B. Boos, and B. V. Shanabrook, "Mobility enhancement in strained p-InGaSb quantum wells," *Applied Physics Letters*, vol. 91, no. 4, p. 042104, 2007.

[77] 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 (V_{cc}=0.5 V) logic applications," in *IEEE International Electron Devices Meeting (IEDM) Technical Digest*, 2008, pp. 1109–1112.

[78] L. Xia, J. B. Boos, B. R. Bennett, M. G. Ancona, and J. A. del Alamo, "Hole mobility enhancement in $In_{0.41}Ga_{0.59}Sb$ quantum-well field-effect transistors," *Applied Physics Letters*, vol. 98, no. 5, p. 053505, 2011.

[79] A. Nainani, S. Raghunathan, D. Witte, M. Kobayashi, T. Irisawa, T. Krishnamohan, K. Saraswat, B. R. Bennett, M. G. Ancona, and J. B. Boos, "Engineering of strained III-V heterostructures for high hole mobility," in *IEEE International Electron Devices Meeting (IEDM) Technical Digest*, 2009, pp. 857–860.

[80] L. Xia, V. Tokranov, S. R. Oktyabrsky, and J. A. del Alamo, "Performance enhancement of p-channel InGaAs quantum-well FETs by superposition of process-induced uniaxial strain and epitaxially-grown biaxial strain," in *IEEE International Electron Devices Meeting (IEDM) Technical Digest*, 2011, pp. 315–318.

[81] P. Nagaiah, V. Tokranov, M. Yakimov, S. Koveshnikov, S. Oktyabrsky, D. Veksler, W. Tsai, and G. Bersuker, "Mobility and remote scattering in buried InGaAs quantum well channels with high-k gate oxide," *Journal of Vacuum Science & Technology B*, vol. 28, no. 3, pp. C3H5–C3H9, 2010.

[82] F. Ren, M. Hong, W. Hobson, J. Kuo, J. Lothian, J. Mannaerts, J. Kwo, S. Chu, Y. Chen, and A. Cho, "Demonstration of enhancement-mode p- and n-channel GaAs MOSFETS with $Ga_2O_3(Gd_2O_3)$As gate oxide," *Solid-State Electronics*, vol. 41, no. 11, pp. 1751–1753, 1997.

[83] M. Passlack, J. K. Abrokwah, R. Droopad, Z. Yu, C. Overgaard, S. I. Yi, M. Hale, J. Sexton, and A. C. Kummel, "Self-aligned GaAs p-channel enhancement mode MOS heterostructure field-effect transistor," *IEEE Electron Device Letters*, vol. 23, no. 9, pp. 508–510, 2002.

[84] M. Xu, R. Wang, and P. Ye, "GaSb inversion-mode pMOSFETs with atomic-layer-deposited Al_2O_3 as gate dielectric," *IEEE Electron Device Letters*, vol. 32, no. 7, pp. 883–885, 2011.

[85] A. Nainani, T. Irisawa, Z. Yuan, B. Bennett, J. Boos, Y. Nishi, and K. Saraswat, "Optimization of the Al_2O_3/GaSb interface and a high-mobility GaSb pMOSFET," *IEEE Transactions on Electron Devices*, vol. 58, no. 10, pp. 3407–3415, 2011.

[86] W. Li-Shu, S. Bing, C. Hu-Dong, Z. Wei, X. Bai-Qing, Z. Xiong, and L. Hong-Gang, "GaSb p-channel metal-oxide-semiconductor field-effect transistors with Ni/Pt/Au source/drain ohmic contacts," *Chinese Physics Letters*, vol. 29, no. 12, p. 127303, 2012.

[87] R. L. Chu, T. H. Chiang, W. J. Hsueh, K. H. Chen, K. Y. Lin, G. J. Brown, J. I. Chyi, J. Kwo, and M. Hong, "Passivation of GaSb using molecular beam epitaxy Y_2O_3 to achieve low interfacial trap density and high-performance self-aligned inversion-channel p-metal-oxide-semiconductor field-effect-transistors," *Applied Physics Letters*, vol. 105, no. 18, p. 182106, 2014.

[88] M. Yokoyama, K. Nishi, S. Kim, H. Yokoyama, M. Takenaka, and S. Takagi, "Self-aligned Ni-GaSb source/drain junctions for GaSb p-channel metal-oxide-semiconductor field-effect transistors," *Applied Physics Letters*, vol. 104, no. 9, p. 093509, 2014.

[89] Z. Yuan, A. Kumar, C.-Y. Chen, A. Nainani, B. Bennett, J. Boos, and K. Saraswat, "Antimonide-based heterostructure p-channel MOSFETs with Ni-alloy source/drain," *Electron Device Letters, IEEE*, vol. 34, no. 11, pp. 1367–1369, 2013.

[90] K. Nishi, M. Yokoyama, H. Yokoyama, T. Hoshi, H. Sugiyama, M. Takenaka, and S. Takagi, "Operation of the GaSb p-channel metal-oxide-semiconductor field-effect transistors fabricated on (111)A surfaces," *Applied Physics Letters*, vol. 105, no. 23, p. 233503, 2014.

[91] P. Nagaiah, V. Tokranov, M. Yakimov, S. Madisetti, A. Greene, S. Novak, R. Moore, H. Bakhru, and S. Oktyabrsky, "In situ deposited HfO_2 with amorphous-Si passivation as a potential gate stack for high mobility (In)GaSb-based p-MOSFETs," *ECS Transactions*, vol. 41, no. 3, pp. 223–230, 2011.

[92] A. Nainani, Z. Yuan, T. Krishnamohan, B. Bennett, J. Boos, M. Reason, M. Ancona, Y. Nishi, and K. Saraswat, "$In_xGa_{1-x}Sb$ channel p-metal-oxide-semiconductor field effect transistors: effect of strain and heterostructure design," *Journal of Applied Physics*, vol. 110, no. 1, p. 014503, 2011.

[93] A. Nainani, T. Irisawa, Z. Yuan, Y. Sun, T. Krishnamohan, M. Reason, B. Bennett, J. Boos, M. Ancona, Y. Nishi, and K. Saraswat, "Development of high-k dielectric for antimonides and a sub 350°C III-V pMOSFET outperforming germanium," in *IEEE International Electron Devices Meeting (IEDM) Technical Digest*, 2010, pp. 138–141.

[94] 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, no. 10, pp. 2454–2459, 2006.

[95] B. Streetman and S. Banerjee, *Solid State Electronic Devices*, ser. Prentice-Hall series in solid state physical electronics. Pearson/Prentice Hall, 2010.

[96] L. Yau, "A simple theory to predict the threshold voltage of short-channel IGFET's," *Solid-State Electronics*, vol. 17, no. 10, pp. 1059–1063, 1974.

[97] E. Nicollian and J. Brews, *MOS (Metal Oxide Semiconductor) Physics and Technology.* New Jersey, NJ: Wiley, 2002, pp. 139, 130; 331.

[98] S. Takagi and A. Toriumi, "Quantitative understanding of inversion-layer capacitance in Si MOSFET's," *IEEE Transactions on Electron Devices*, vol. 42, no. 12, pp. 2125–2130, 1995.

[99] P. K. Hurley, R. Long, T. O'Regan, E. O'Connor, S. Monaghan, V. Djara, M. A. Negara, A. O'Mahony, I. Povey, A. Blake, R. Nagle, D. O'Connell, M. Pemble, and K. Cherkaoui, "(Invited) equivalent oxide thickness correction in the high-k/$In_{0.53}Ga_{0.47}As$/InP system," *ECS Transactions*, vol. 33, no. 3, pp. 433–444, 2010.

[100] S. Luryi, "Quantum capacitance devices," *Applied Physics Letters*, vol. 52, no. 6, pp. 501–503, 1988.

[101] G. Brammertz, K. Martens, S. Sioncke, A. Delabie, M. Caymax, M. Meuris, and M. Heyns, "Characteristic trapping lifetime and capacitance-voltage measurements of GaAs metal-oxide-semiconductor structures," *Applied Physics Letters*, vol. 91, no. 13, p. 133510, 2007.

[102] E. J. Kim, L. Wang, P. M. Asbeck, K. C. Saraswat, and P. C. McIntyre, "Border traps in Al_2O_3/$In_{0.53}Ga_{0.47}As$ (100) gate stacks and their passivation by hydrogen anneals," *Applied Physics Letters*, vol. 96, no. 1, p. 012906, 2010.

[103] Y. Yuan, L. Wang, B. Yu, B. Shin, J. Ahn, P. C. McIntyre, P. Asbeck, M. J. W. Rodwell, and Y. Taur, "A distributed model for border traps in Al_2O_3-InGaAs MOS devices," *IEEE Electron Device Letters*, vol. 32, no. 4, pp. 485–487, 2011.

[104] N. Bhat and K. C. Saraswat, "Characterization of border trap generation in rapid thermally annealed oxides deposited using silane chemistry," *Journal of Applied Physics*, vol. 84, no. 5, pp. 2722–2726, 1998.

[105] P. C. Mathur, R. Shyam, and S. Jain, "Transport phenomena in III-V compound semiconductors," *physica status solidi (a)*, vol. 50, no. 1, pp. 11–40, 1978.

[106] J. J. Harris, J. A. Pals, and R. Woltjer, "Electronic transport in low-dimensional structures," *Reports on Progress in Physics*, vol. 52, no. 10, pp. 1217–1265, 1989.

[107] H. L. Stőrmer, K. Baldwin, A. C. Gossard, and W. Wiegmann, "Modulation-doped field-effect transistor based on a two-dimensional hole gas," *Applied Physics Letters*, vol. 44, no. 11, pp. 1062–1064, 1984.

[108] D. J. Paul, "Si/SiGe heterostructures: from material and physics to devices and circuits," *Semiconductor Science and Technology*, vol. 19, no. 10, pp. R75–R108, 2004.

[109] R. People and J. C. Bean, "Calculation of critical layer thickness versus lattice mismatch for Ge_xSi_{1-x}/Si strained-layer heterostructures," *Applied Physics Letters*, vol. 47, no. 3, pp. 322–324, 1985.

[110] J. Matthews and A. Blakeslee, "Defects in epitaxial multilayers: I. misfit dislocations," *Journal of Crystal Growth*, vol. 27, pp. 118–125, 1974.

[111] S. Oktyabrsky, "p-type channel field-effect transistors," in *Fundamentals of III-V Semiconductor MOSFETs*, S. Oktyabrsky and P. Ye, Eds. Springer US, 2010, pp. 349–378.

[112] S. Chou and D. Antoniadis, "Relationship between measured and intrinsic transconductances of FET's," *Electron Devices, IEEE Transactions on*, vol. 34, no. 2, pp. 448–450, 1987.

[113] A. Ali, H. Madan, R. Misra, E. Hwang, A. Agrawal, I. Ramirez, P. Schiffer, T. Jackson, S. E. Mohney, J. B. Boos, B. Bennett, I. Geppert, M. Eizenberg, and S. Datta, "Advanced composite high-k gate stack for mixed anion arsenide-antimonide quantum well transistors," in *IEEE International Electron Devices Meeting (IEDM) Technical Digest*, 2010, pp. 134–137.

[114] W. Schottky, "Vereinfachte und erweiterte theorie der randschicht-gleichrichter," *Zeitschrift für Physik*, vol. 118, no. 9, pp. 539–592.

[115] J. Bardeen, "Surface states and rectification at a metal semi-conductor contact," *Phys. Rev.*, vol. 71, pp. 717–727, 1947.

[116] M. Murakami, "Development of refractory ohmic contact materials for gallium arsenide compound semiconductors," *Science and Technology of Advanced Materials*, vol. 3, no. 1, pp. 1–27, 2002.

[117] E. Reichmanis and L. F. Thompson, "Polymer materials for microlithography," *Chemical Reviews*, vol. 89, no. 6, pp. 1273–1289, 1989.

[118] D. Mailly and C. Vieu, "Lithography and etching processes," in *Nanoscience*, C. Dupas, P. Houdy, and M. Lahmani, Eds. Springer Berlin Heidelberg, 2007, pp. 3–40.

[119] M. Rothschild, T. M. Bloomstein, N. Efremow, T. H. Fedynyshyn, M. Fritze, I. Pottebaum, and M. Switkes, "Nanopatterning with UV optical lithography," *MRS Bulletin*, vol. 30, pp. 942–946, 2005.

[120] T. A. Brunner, "Why optical lithography will live forever," *Journal of Vacuum Science & Technology B: Microelectronics and Nanometer Structures*, vol. 21, no. 6, pp. 2632–2637, 2003.

[121] S. P. Beaumont, P. G. Bower, T. Tamamura, and C. D. W. Wilkinson, "Sub-20-nm-wide metal lines by electron-beam exposure of thin poly(methyl methacrylate) films and liftoff," *Applied Physics Letters*, vol. 38, no. 6, pp. 436–439, 1981.

[122] Z. Cui and Z. Cui, "Nanofabrication by charged beams," in *Nanofabrication*. Springer US, 2008, pp. 77–127.

[123] C. S. Wu, Y. Makiuchi, and C. Chen, "High-energy electron beam lithography for nanoscale fabrication," in *Lithography*, M. Wang, Ed. InTech, 2010, pp. 241–266.

[124] J. N. Helbert, "Electron-beam ULSI applications," in *Handbook of VLSI Microlithography*, W. Andrew, Ed. London, 2001, pp. 670–755.

[125] L. Ren and B. Chen, "Proximity effect in electron beam lithography," in *International Conference on Solid-State and Integrated Circuits Technology Proceedings*, vol. 1, 2004, pp. 579–582.

[126] T. H. P. Chang, "Proximity effect in electron beam lithography," *Journal of Vacuum Science and Technology*, vol. 12, no. 6, pp. 1271–1275, 1975.

[127] M. Mohammad, M. Muhammad, S. Dew, and M. Stepanova, "Fundamentals of electron beam exposure and development," in *Nanofabrication*, M. Stepanova and S. Dew, Eds. Springer Vienna, 2012, pp. 11–41.

[128] S. Yasin, D. Hasko, and H. Ahmed, "Comparison of MIBK/IPA and water/IPA as PMMA developers for electron beam nanolithography," *Microelectronic Engineering*, vol. 6162, p. 745753, 2002.

[129] D. E. Davis, "Registration mark detection for electron-beam lithography-EL1 system," *IBM Journal of Research and Development*, vol. 24, no. 5, pp. 545–553, 1980.

[130] 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, no. 56, pp. 761–763, 2008.

[131] A. Stamper, T. McDevitt, and S. Luce, "Sub-0.25-micron interconnection scaling: damascene copper versus subtractive aluminum," in *IEEE/SEMI Advanced Semiconductor Manufacturing Conference and Workshop*, 1998, pp. 337–346.

[132] S. Murad, M. Rahman, N. Johnson, S. Thoms, S. P. Beaumont, and C. D. W. Wilkinson, "Dry etching damage in III-V semiconductors," *Journal of Vacuum Science Technology B: Microelectronics and Nanometer Structures*, vol. 14, no. 6, pp. 3658–3662, 1996.

[133] M. J. Rooks, S. Wind, P. McEuen, and D. E. Prober, "Fabrication of 30-nm-scale structures for electron transport studies using a polymethylmethacrylate bilayer resist," *Journal of Vacuum Science & Technology B: Microelectronics and Nanometer Structures*, vol. 5, no. 1, pp. 318–321, 1987.

[134] J. Golden, H. Miller, and D. N. J. Ross, "Optimization of bi-layer lift-off resist process," *CS Mantech Technical Digest*, 2009.

[135] I. T. Ferguson, A. G. Thompson, S. A. Barnett, F. H. Long, and Z. C. Feng, "Epitaxial film growth and characterization," in *Frontiers of Thin Film Technology*, ser. Thin Films, M. H. Francombe, Ed. Elsevier, 2001, vol. 28, pp. 1–69.

[136] C. A. Zorman, R. C. Roberts, and L. Chen, "Additive processes for semiconductors and dielectric materials," in *MEMS Materials and Processes Handbook*, ser. MEMS Reference Shelf, R. Ghodssi and P. Lin, Eds. Springer US, 2011, pp. 37–136.

[137] H. Zhou, C. Sim, A. Glidle, C. Hodson, C. R. Kinsey, and C. D. W. Wilkinson, *Properties of Silicon Nitride by Room-Temperature Inductively Coupled Plasma Deposition.* Wiley-VCH Verlag GmbH & Co. KGaA, 2005, pp. 77–86.

[138] R. Williams, *Modern GaAs Processing Methods.* Boston: Artech House, 1990, pp. 153–155.

[139] J. W. Lee, K. D. Mackenzie, D. Johnson, J. N. Sasserath, S. J. Pearton, and F. Ren, "Low temperature silicon nitride and silicon dioxide film processing by inductively coupled plasma chemical vapor deposition," *Journal of The Electrochemical Society*, vol. 147, no. 4, pp. 1481–1486, 2000.

[140] J. Kim and T. Kim, "Initial surface reactions of atomic layer deposition," *MRS Proceedings*, vol. 61, pp. 17–22, 2009.

[141] A. Foroughi-Abari and K. Cadien, "Atomic layer deposition for nanotechnology," in *Nanofabrication*, M. Stepanova and S. Dew, Eds. Springer Vienna, 2012, pp. 143–161.

[142] S. M. George, "Atomic layer deposition: an overview," *Chemical Reviews*, vol. 110, no. 1, pp. 111–131, 2010.

[143] C.-H. Chen, E. L. Hu, W. V. Schoenfeld, and P. M. Petroff, "Metallization-induced damage in III-V semiconductors," *Journal of Vacuum Science Technology B: Microelectronics and Nanometer Structures*, vol. 16, no. 6, pp. 3354–3358, 1998.

[144] A. Baca and C. Ashby, *Fabrication of GaAs Devices*. Philadephia, PA: Institution of Engineering and Technology, 2005, pp. 117–145.

[145] M. Nawaz, W. Strupinski, J. Stenarson, S. Persson, and H. Zirath, "MOCVD grown AlInAs/-GaInAs/InP HEMTs with InP etch-stop layer," in *International Conference on Microwaves, Radar and Wireless Communications Proceedings*, vol. 1, 2000, pp. 116–120.

[146] A. Baca and C. Ashby, *Fabrication of GaAs Devices*. Philadephia, PA: Institution of Engineering and Technology, 2005, pp. 147–178.

[147] D. Schroder, "Mobility," in *Semiconductor Material and Device Characterization*, 3rd ed. John Wiley & Sons, Inc., 2005, pp. 465–522.

[148] I. Isenberg, B. R. Russell, and R. F. Greene, "Improved method for measuring Hall coefficients," *Review of Scientific Instruments*, vol. 19, no. 10, pp. 685–688, 1948.

[149] D. C. Look, "The Hall effect and magnetoresistance," in *Electrical Characterization of GaAs Materials and Devices*, 3rd ed. John Wiley & Sons, Inc., 1991, pp. 1–131.

[150] S. Jandl, K. D. Usadel, and G. Fischer, "Resistivity measurements with samples in the form of a double cross," *Review of Scientific Instruments*, vol. 45, no. 11, pp. 1479–1480, 1974.

[151] 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.

[152] A. Goetzberger, R. Scarlett, and W. Shockley, "Research and investigation of inverse epitaxial UHF power transistors," oai.dtic.mil, 1964.

[153] D. Sawdai, D. Pavlidis, and D. Cui, "Enhanced transmission line model structures for accurate resistance evaluation of small-size contacts and for more reliable fabrication," *IEEE Transactions on Electron Devices*, vol. 46, no. 7, pp. 1302–1311, 1999.

[154] H. Murrmann and D. Widmann, "Current crowding on metal contacts to planar devices," *Electron Devices, IEEE Transactions on*, vol. 16, no. 12, pp. 1022–1024, 1969.

[155] G. Reeves and H. B. Harrison, "Obtaining the specific contact resistance from transmission line model measurements," *IEEE Electron Device Letters*, vol. 3, no. 5, pp. 111–113, 1982.

[156] D. Schroder, "Oxide and interface trapped charges, oxide thickness," in *Semiconductor Material and Device Characterization*, 3rd ed. John Wiley & Sons, Inc., 2005, pp. 319–388.

[157] R. D. Long, B. Shin, S. Monaghan, K. Cherkaoui, J. Cagnon, S. Stemmer, P. C. McIntyre, and P. K. Hurley, "Charged defect quantification in $Pt/Al_2O_3/In_{0.53}Ga_{0.47}As/InP$ MOS capacitors," *Journal of The Electrochemical Society*, vol. 158, no. 5, pp. G103–G107, 2011.

[158] R. Hillard, J. Heddleson, D. Zier, P. Rai-Choudhury, and D. Schroder, *Direct and rapid method for determining flatband voltage from non-equilibrium capacitance voltage data*. Pennington, NJ, 1992, pp. 261–274.

[159] E. H. Nicollian and A. Goetzberger, "The Si-SiO_2 interface electrical properties as determined by the metal-insulator-silicon conductance technique," *Bell System Technical Journal*, vol. 46, no. 6, pp. 1055–1133, 1967.

[160] R. Engel-Herbert, Y. Hwang, and S. Stemmer, "Comparison of methods to quantify interface trap densities at dielectric/III-V semiconductor interfaces," *Journal of Applied Physics*, vol. 108, no. 12, p. 124101, 2010.

[161] K. Martens, C. O. Chui, G. Brammertz, B. De Jaeger, D. Kuzum, M. Meuris, M. M. Heyns, T. Krishnamohan, K. Saraswat, H. Maes, and G. Groeseneken, "On the correct extraction of interface trap density of MOS devices with high-mobility semiconductor substrates," *Electron Devices, IEEE Transactions on*, vol. 55, no. 2, pp. 547–556, 2008.

[162] H. C. Lin, G. Brammertz, K. Martens, G. de Valicourt, L. Negre, W.-E. Wang, W. Tsai, M. Meuris, and M. Heyns, "The Fermi-level efficiency method and its applications on high interface trap density oxide-semiconductor interfaces," *Applied Physics Letters*, vol. 94, no. 15, p. 153508, 2009.

[163] K. Martens, W. Wang, K. D. Keersmaecker, G. Borghs, G. Groeseneken, and H. Maes, "Impact of weak Fermi-level pinning on the correct interpretation of III-V MOS C-V and G-V characteristics," *Microelectronic Engineering*, vol. 84, no. 9-10, pp. 2146–2149, 2007.

[164] L. Terman, "An investigation of surface states at a silicon/silicon oxide interface employing metal-oxide-silicon diodes," *Solid-State Electronics*, vol. 5, no. 5, pp. 285–299, 1962.

[165] C. N. Berglund, "Surface states at steam-grown silicon-silicon dioxide interfaces," *IEEE Transactions on Electron Devices*, vol. 13, no. 10, pp. 701–705, 1966.

[166] R. Castagné and A. Vapaille, "Description of the SiO_2-Si interface properties by means of very low frequency MOS capacitance measurements," *Surface Science*, vol. 28, no. 1, pp. 157–193, 1971.

[167] D. Mui, Z. Wang, and H. Morko, "A review of IIIV semiconductor based metal-insulator-semiconductor structures and devices," *Thin Solid Films*, vol. 231, no. 12, pp. 107–124, 1993.

[168] A. Ortiz-Conde, F. G. Snchez, J. Liou, A. Cerdeira, M. Estrada, and Y. Yue, "A review of recent MOSFET threshold voltage extraction methods," *Microelectronics Reliability*, vol. 42, no. 45, pp. 583–596, 2002.

[169] D.-H. Kim, T.-W. Kim, R. Hill, C. Young, C. Y. Kang, C. Hobbs, P. Kirsch, J. del Alamo, and R. Jammy, "High-speed e-mode InAs QW MOSFETs With Al_2O_3 insulator for future RF applications," *Electron Device Letters, IEEE*, vol. 34, no. 2, pp. 196–198, 2013.

[170] S. Lee, C.-Y. Huang, D. Cohen-Elias, B. Thibeault, W. Mitchell, V. Chobpattana, S. Stemmer, A. Gossard, and M. Rodwell, "Highly scalable raised source/drain InAs quantum well MOSFETs exhibiting I_{ON} = 482 μ A/μ m at I_{OFF} = 100 nA/μ m and V_{DD} = 0.5V," *Electron Device Letters, IEEE*, vol. 35, no. 6, pp. 621–623, 2014.

[171] J. Koomen, "Investigation of the MOST channel conductance in weak inversion," *Solid-State Electronics*, vol. 16, no. 7, pp. 801–810, 1973.

[172] C. Sodini, T. Ekstedt, and J. Moll, "Charge accumulation and mobility in thin dielectric MOS transistors," *Solid-State Electronics*, vol. 25, no. 9, pp. 833–841, 1982.

[173] G. W. Paterson, S. J. Bentley, M. C. Holland, I. G. Thayne, J. Ahn, R. D. Long, P. C. McIntyre, and A. R. Long, "Admittance and subthreshold characteristics of atomic-layer-deposition Al_2O_3 on $In_{0.53}Ga_{0.47}As$ in surface and buried channel flatband metal-oxide-semiconductor field effect transistors," *Journal of Applied Physics*, vol. 111, no. 10, p. 104112, 2012.

[174] R. Van Overstraeten, G. Declerck, and P. Muls, "Theory of the MOS transistor in weak inversion - new method to determine the number of surface states," *Electron Devices, IEEE Transactions on*, vol. 22, no. 5, pp. 282–288, 1975.

[175] P. Bai, C. Auth, S. Balakrishnan, M. Bost, R. Brain, V. Chikarmane, R. Heussner, M. Hussein, J. Hwang, D. Ingerly, R. James, J. Jeong, C. Kenyon, E. Lee, S.-H. Lee, N. Lindert, M. Liu, Z. Ma, T. Marieb, A. Murthy, R. Nagisetty, S. Natarajan, J. Neirynck, A. Ott, C. Parker, J. Sebastian, R. Shaheed, S. Sivakumar, J. Steigerwald, S. Tyagi, C. Weber, B. Woolery, A. Yeoh, K. Zhang, and M. Bohr, "A 65nm logic technology featuring 35 nm gate lengths, enhanced channel strain, 8 Cu interconnect layers, low-k ILD and 0.57 μm^2 SRAM cell," in *IEEE International Electron Devices Meeting (IEDM) Technical Digest*, 2004, pp. 657–660.

[176] Y. Zhang and M. Fischetti, "Calculation of hole mobility in Ge and III-V p-channels," in *Computational Electronics, 2009. IWCE '09. 13th International Workshop on*, May 2009, pp. 1–4.

[177] B. R. Bennett, M. G. Ancona, and J. B. Boos, "Compound semiconductors for low-power p-channel field-effect transistors," *MRS Bulletin*, vol. 34, no. 7, pp. 530–536, 2009.

[178] L. Nguyen, L. Larson, and U. K. Mishra, "Ultra-high speed modulation-doped field-effect transistors: a tutorial review," *Proceedings of the IEEE*, vol. 80, no. 4, pp. 494–518, 1992.

[179] I. Thayne, R. Hill, M. Holland, X. Li, H. Zhou, D. Macintyre, S. Thoms, K. Kalna, C. Stanley, A. Asenov, R. Droopad, and M. Passlack, "Review of current status of III-V MOSFETs," *ECS Transactions*, vol. 19, no. 5, pp. 275–286, 2009.

[180] S. Bentley, M. Holland, X. Li, G. Paterson, H. Zhou, O. Ignatova, D. Macintyre, S. Thoms, A. Asenov, B. Shin, J. Ahn, P. McIntyre, and I. Thayne, "Electron mobility in surface- and buried-channel flatband $In_{0.53}Ga_{0.47}As$ MOSFETs with ALD Al_2O_3 gate dielectric," *IEEE Electron Device Letters*, vol. 32, no. 4, pp. 494–496, 2011.

[181] H. Zhao, Y.-T. Chen, J. H. Yum, Y. Wang, F. Zhou, F. Xue, and J. C. Lee, "Effects of barrier layers on device performance of high mobility $In_{0.7}Ga_{0.3}As$ metal-oxide-semiconductor field-effect-transistors," *Applied Physics Letters*, vol. 96, no. 10, 2010.

[182] 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/V.s, and transconductance of over 475 $\mu S/\mu m$," *IEEE Electron Device Letters*, vol. 28, no. 12, pp. 1080–1082, 2007.

[183] Y. Sun, E. Kiewra, J. de Souza, J. Bucchignano, K. Fogel, D. Sadana, and G. Shahidi, "High performance long- and short-Channel $In_{0.7}Ga_{0.3}As$-channel MOSFETs," in *Device Research Conference (DRC)*, 2008, pp. 41–42.

[184] J. Huang, N. Goel, H. Zhao, C. Kang, K. Min, G. Bersuker, S. Oktyabrsky, C. Gaspe, M. Santos, P. Majhi, P. Kirsch, H.-H. Tseng, J. Lee, and R. Jammy, "InGaAs MOSFET performance and reliability improvement by simultaneous reduction of oxide and interface charge in ALD (La)AlOx/ZrO_2 gate stack," in *IEEE International Electron Devices Meeting (IEDM) Technical Digest*, 2009, pp. 335–338.

[185] D. Lin, N. Waldron, G. Brammertz, K. Martens, W.-E. Wang, S. Sioncke, A. Delabie, H. Bender, T. Conard, W. H. Tseng, J. Lin, K. Temst, A. Vantomme, J. Mitard, M. Caymax, M. Meuris, M. Heyns, and T. Hoffman, "(Invited) exploring the ALD $Al_2O_3/In_{0.53}Ga_{0.47}As$ and Al_2O_3/Ge interface properties: a common gate stack approach for advanced III-V/Ge CMOS," *ECS Transactions*, vol. 28, no. 5, pp. 173–183, 2010.

[186] A. Ali, H. Madan, S. Koveshnikov, S. Oktyabrsky, R. Kambhampati, T. Heeg, D. Schlom, and S. Datta, "Small-signal response of inversion layers in high-mobility $In_{0.53}Ga_{0.47}As$ MOSFETs made with thin high-k dielectrics," *IEEE Transactions on Electron Devices*, vol. 57, no. 4, pp. 742–748, 2010.

[187] R. T. Hsu, W. C. Hsu, J. S. Wang, M. J. Kao, Y. H. Wu, and J. S. Su, "Study of zinc-delta-doped strained quantum well InGaAs/GaAs p-channel heterostructure field-effect transistors," *Japanese Journal of Applied Physics*, vol. 35, no. Part 1, No. 4A, pp. 2085–2089, 1996.

[188] J. Pearson, M. Holland, C. Stanley, A. Long, E. Skuras, A. Asenov, and J. Davies, "Optimization of layer structure for InGaAs channel pseudomorphic HEMTs," *Journal of Crystal Growth*, vol. 201–202, pp. 757–760, 1999.

[189] A. Carvalho, A. De Oliveira, M. Moreira, and E. Alves, "Migration of silicon atoms in planar-doped GaAs/AlGaAs modulation doped fluid effect transistor heterostructures grown by molecular beam epitaxy," *Materials Science & Engineering B*, vol. 35, no. 1-3, pp. 149–155, 1995.

[190] M. Hirano, K. Oe, and F. Yanagawa, "p-channel AlGaAs/GaAs heterostructure FETs employing two-dimensional hole gas," *Japanese Journal of Applied Physics*, vol. 23, no. Part 2, No. 11, pp. L868–L870, 1984.

[191] S. Tiwari and W. Wang, "p-Channel MODFET's using GaAlAs/GaAs two-dimensional hole gas," *IEEE Electron Device Letters*, vol. 5, no. 8, pp. 333–335, 1984.

[192] M. Kudo, H. Matsumoto, T. Tanimoto, T. Mishima, and I. Ohbu, "Improved hole transport properties of highly strained $In_{0.35}Ga_{0.65}As$ channel double-modulation-doped structures grown by MBE on GaAs," *Journal of Crystal Growth*, vol. 175-176, Part 2, pp. 910–914, 1997.

[193] T. J. Drummond, T. E. Zipperian, I. J. Fritz, J. E. Schirber, and T. A. Plut, "p-channel, strained quantum well, field-effect transistor," *Applied Physics Letters*, vol. 49, no. 8, pp. 461–463, 1986.

[194] C.-P. Lee, H. Wang, G. Sullivan, N. Sheng, and D. Miller, "High-transconductance p-channel InGaAs/AlGaAs modulation-doped field effect transistors," *IEEE Electron Device Letters*, vol. 8, no. 3, pp. 85–87, 1987.

[195] P. Ruden, M. Shur, D. Arch, R. Daniels, D. Grider, and T. Nohava, "Quantum-well p-channel AlGaAs/InGaAs/GaAs heterostructure insulated-gate field-effect transistors," *IEEE Transactions on Electron Devices*, vol. 36, no. 11, pp. 2371–2379, 1989.

[196] M. Kudo, T. Mishima, H. Matsumoto, I. Ohbu, and T. Tanimoto, "Highly strained $\ln_{0.35}Ga_{0.65}As$/GaAs layers grown by molecular beam epitaxy for high hole mobility transistors," *Journal of Electronic Materials*, vol. 25, pp. 944–947, 1996.

[197] J. Chen, A. Li, Q. Yang, C. Lin, Y. Ren, S. Jin, M. Qi, H. Xu, and X. Chen, "GSMBE grown $In_{0.49}Ga_{0.51}P$/(In)GaAs/GaAs high hole mobility transistor structures," *Journal of Crystal Growth*, vol. 201–202, pp. 744–748, 1999.

[198] H. J. Kim, D. M. Kim, D. H. Woo, S.-I. Kim, S. H. Kim, J. I. Lee, K. N. Kang, and K. Cho, "High photoresponsivity of a p-channel InGaP/GaAs/InGaAs double heterojunction pseudomorphic modulation-doped field effect transistor," *Applied Physics Letters*, vol. 72, no. 5, pp. 584–586, 1998.

[199] T. Tanimoto, M. Yamane, S. Goto, and Y. Uchida, "An analytical simulation of p-AlGaAs/InGaAs/GaAs pseudomorphic FET," *IEICE Transactions*, vol. E73, no. 11, pp. 1849–1851, 1990.

[200] R. T. Hsu, W. C. Hsu, M. J. Kao, and J. S. Wang, "Characteristics of a delta-doped GaAs/InGaAs p-channel heterostructure field-effect transistor," *Applied Physics Letters*, vol. 66, no. 21, pp. 2864–2866, 1995.

[201] A. Kusters, A. Kohl, V. Sommer, R. Muller, and K. Heime, "Optimized double heterojunction pseudomorphic InP/$In_xGa_{1-x}As$/InP ($0.64 \leq x \leq 0.82$) p-MODFETs and the role of strain in their design," *IEEE Transactions on Electron Devices*, vol. 40, no. 12, pp. 2164–2170, 1993.

[202] Y.-J. Chan and D. Pavlidis, "Single and dual p-doped channel $In_{0.52}Al_{0.48}As$/$In_xGa_{1-x}As$ (x=0.53, 0.65) FET's and the role of doping," *IEEE Transactions on Electron Devices*, vol. 39, no. 3, pp. 466–472, 1992.

[203] L. Luo, K. Longenbach, and W. Wang, "p-channel modulation-doped field-effect transistors based on $AlSb_{0.9}As_{0.1}$/GaSb," *IEEE Electron Device Letters*, vol. 11, no. 12, pp. 567–569, 1990.

[204] J. Klem, J. Lott, J. Schirber, S. Kurtz, and S. Lin, "Strained quantum well modulation-doped InGaSb/AlGaSb structures grown by molecular beam epitaxy," *Journal of Electronic Materials*, vol. 22, pp. 315–318, 1993.

[205] J. B. Boos, B. R. Bennett, N. A. Papanicolaou, M. G. Ancona, J. G. Champlain, Y.-C. Chou, M. D. Lange, J. M. Yang, R. Bass, D. Park, and B. V. Shanabrook, "Sb-based n- and p-channel heterostructure FETs for high-speed, low-power applications," *IEICE Transactions on Electronics*, vol. E91.C, no. 7, pp. 1050–1057, 2008.

[206] M. Kobayashi, G. Thareja, Y. Sun, N. Goel, M. Garner, W. Tsai, P. Pianetta, and Y. Nishi, "The effects of wet surface clean and in situ interlayer on $In_{0.52}Al_{0.48}As$ metal-oxide-semiconductor characteristics," *Applied Physics Letters*, vol. 96, no. 14, p. 142906, 2010.

[207] T. Shibata, J. Nakata, Y. Nanishi, and M. Fujimoto, "A rutherford backscattering spectroscopic study of the aluminum antimonide oxidation process in air," *Japanese Journal of Applied Physics*, vol. 33, no. Part 1, No. 4A, pp. 1767–1772, 1994.

[208] A. Rosenberg, "The oxidation of intermetallic compounds-III: the room-temperature oxidation of $A^{III}B^{V}$ compounds," *Journal of Physics and Chemistry of Solids*, vol. 14, pp. 175–180, 1960.

[209] A. Nainani, Y. Sun, T. Irisawa, Z. Yuan, M. Kobayashi, P. Pianetta, B. R. Bennett, J. B. Boos, and K. C. Saraswat, "Device quality Sb-based compound semiconductor surface: a comparative study of chemical cleaning," *Journal of Applied Physics*, vol. 109, no. 11, p. 114908, 2011.

[210] S. J. Pearton, W. S. Hobson, A. E. Von Neida, N. M. Haegel, K. S. Jones, N. Morris, and B. J. Sealy, "Implant activation and redistribution in $Al_xGa_{1-x}As$," *Journal of Applied Physics*, vol. 67, no. 5, pp. 2396–2409, 1990.

[211] C. C. Lee, M. D. Deal, and J. C. Bravman, "Diffusion of implanted Be in $Al_xGa_{1-x}As$ as a function of Al concentration and anneal temperature," *Applied Physics Letters*, vol. 66, no. 3, pp. 355–357, 1995.

[212] J. J. Harris, "Simplified assessment of parallel conduction in modulation-doped heterostructures," *Measurement Science and Technology*, vol. 2, no. 12, pp. 1201–1205, 1991.

[213] S. Battersby, F. Selten, J. Harris, and C. Foxon, "Magnetoresistance effect in AlGaAs/-GaAs two-dimensional electron gas structures at room temperature," *Solid-State Electronics*, vol. 31, no. 6, pp. 1083–1088, 1988.

[214] I. J. Fritz, L. R. Dawson, T. J. Drummond, J. E. Schirber, and R. M. Biefeld, "Hall-effect measurements in *p*-type InGaAs/GaAs strained-layer superlattices," *Applied Physics Letters*, vol. 48, no. 2, pp. 139–141, 1986.

[215] J. J. Harris, J. M. Lagemaat, S. J. Battersby, C. M. Hellon, C. T. Foxon, and D. E. Lacklison, "Sub-band populations and the spatial distribution of electrons in GaAs/(Al,Ga)As modulation-doped quantum wells," *Semiconductor Science and Technology*, vol. 3, no. 8, pp. 773–780, 1988.

[216] T. Mishima, C. W. Fredriksz, G. F. A. van de Walle, D. J. Gravesteijn, R. A. van den Heuvel, and A. A. van Gorkum, "Effect of interface quality on the electrical properties of *p*-Si/SiGe two-dimensional hole gas systems," *Applied Physics Letters*, vol. 57, no. 24, pp. 2567–2569, 1990.

[217] M. A. Sadeghzadeh, C. P. Parry, P. J. Phillips, E. H. C. Parker, and T. E. Whall, "Issues on the molecular-beam epitaxial growth of *p*-SiGe inverted-modulation-doped structures," *Applied Physics Letters*, vol. 74, no. 4, pp. 579–581, 1999.

[218] A. S. Brown, L. D. Nguyen, R. A. Metzger, A. E. Schmitz, and J. A. Henige, "Growth and properties of high mobility strained inverted AlInAs-GaInAs modulation doped structures," *Journal of Vacuum Science & Technology B*, vol. 10, no. 2, pp. 1017–1019, 1992.

[219] L. Pfeiffer, E. F. Schubert, K. W. West, and C. W. Magee, "Si dopant migration and the AlGaAs/GaAs inverted interface," *Applied Physics Letters*, vol. 58, no. 20, pp. 2258–2260, 1991.

[220] J. Lott, J. Klem, and J. Wendt, "Strained p-channel InGaSb/AlGaSb modulation-doped field-effect transistors," *IEEE Electronics Letters*, vol. 28, no. 15, pp. 1459–1460, 1992.

[221] J. Boos, B. Bennett, N. Papanicolaou, M. Ancona, J. Champlain, R. Bass, and B. Shanabrook, "High mobility p-channel HFETs using strained Sb-based materials," *IEEE Electronics Letters*, vol. 43, no. 15, pp. 834–835, 2007.

[222] M. V. Hove, G. Zou, W. D. Raedt, P. Jansen, R. Jonckheere, M. V. Rossum, A. C. F. Hoole, D. R. Allee, A. N. Broers, P. Crozat, Y. Jin, F. Aniel, and R. Adde, "Scaling behavior of delta-doped AlGaAs/InGaAs high electron mobility transistors with gatelengths down to 60 nm and source-drain gaps down to 230 nm," *Journal of Vacuum Science & Technology B: Microelectronics and Nanometer Structures*, vol. 11, no. 4, pp. 1203–1208, 1993.

[223] H. Hasegawa and M. Akazawa, "Interface models and processing technologies for surface passivation and interface control in III-V semiconductor nanoelectronics," *Applied Surface Science*, vol. 254, no. 24, pp. 8005–8015, 2008.

[224] D. A. J. Moran, R. J. W. Hill, X. Li, H. Zhou, D. Mclntyre, S. Thoms, R. Droopad, P. Zurcher, K. Rajagopalan, J. Abrokwah, M. Passlack, and I. Thayne, "Sub-micron, metal gate, high-k dielectric, implant-free, enhancement-mode III-V MOSFETs," in *European Solid State Device Research Conference (ESSDERC) Proceedings*, 2007, pp. 466–469.

[225] E. Kamińska, A. Piotrowska, E. Mizera, and E. Dynowska, "Evolution of the microstructure of Au(Zn) metallization during formation of an ohmic contact to p-type GaAs," *Thin Solid Films*, vol. 246, no. 1-2, pp. 143–150, 1994.

[226] A. Piotrowska, E. Kamińska, S. Kwiatkowski, and A. Turos, "Quantitative analysis of arsenic losses during the formation of Au(Zn)/p-GaAs ohmic contacts," *Journal of Applied Physics*, vol. 73, no. 9, pp. 4404–4408, 1993.

[227] B. Grbić, R. Leturcq, T. Ihn, K. Ensslin, G. Blatter, D. Reuter, and A. D. Wieck, "Hysteretic magnetotransport in p-type AlGaAs heterostructures with In/Zn/Au ohmic contacts," *Physical Review B*, vol. 77, p. 245307, 2008.

[228] A. Barcz, E. Kamińska, and A. Piotrowska, "Fundamental and practical aspects of alloying encapsulated gold-based contacts to GaAs," *Thin Solid Films*, vol. 149, no. 2, pp. 251–260, 1987.

[229] A. Piotrowska, E. Kamińska, A. Barcz, J. Adamczewska, and A. Turos, "Gold-based ohmic contacts on III-V compounds: thermally induced reactions between metallization and the semiconductor compound," *Thin Solid Films*, vol. 130, no. 3-4, pp. 231–236, 1985.

[230] V. Malina, U. Schade, and K. Vogel, "Technological aspects of the preparation of Au-Zn ohmic contacts to p-type InP," *Semiconductor Science and Technology*, vol. 9, no. 1, p. 49, 1994.

[231] Y. Lu, T. S. Kalkur, and C. A. Paz de Araujo, "Rapid thermal alloyed ohmic contacts to p-type GaAs," *Journal of The Electrochemical Society*, vol. 136, no. 10, pp. 3123–3129, 1989.

[232] M. Hirano and F. Yanagawa, "Low-resistance ohmic contacts to p-GaAs," *Japanese Journal of Applied Physics*, vol. 25, no. Part 1, No. 8, pp. 1268–1269, 1986.

[233] Y.-C. Yeo, T.-J. King, and C. Hu, "Metal-dielectric band alignment and its implications for metal gate complementary metal-oxide-semiconductor technology," *Journal of Applied Physics*, vol. 92, no. 12, pp. 7266–7271, 2002.

[234] G. J. Burek, Y. Hwang, A. D. Carter, V. Chobpattana, J. J. M. Law, W. J. Mitchell, B. Thibeault, S. Stemmer, and M. J. W. Rodwell, "Influence of gate metallization processes on the electrical characteristics of high-k/$In_{0.53}Ga_{0.47}As$ interfaces," *Journal of Vacuum Science & Technology B: Microelectronics and Nanometer Structures*, vol. 29, no. 4, p. 040603, 2011.

[235] H. C. Shin and C. Hu, "Thin gate oxide damage due to plasma processing," *Semiconductor Science and Technology*, vol. 11, no. 4, p. 463, 1996.

[236] B. O'Sullivan, L. Pantisano, P. Roussel, R. Degraeve, G. Groeseneken, S. Degendt, and M. Heyns, "On the recovery of simulated plasma process induced damage in high-k dielectrics," in *IEEE International Reliability Physics Symposium Proceedings*, 2006, pp. 365–369.

[237] A. Martin, "Review on the reliability characterization of plasma-induced damage," *Journal of Vacuum Science & Technology B*, vol. 27, no. 1, pp. 426–434, 2009.

[238] D. A. J. 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, no. 12, pp. 2920–2925, 2006.

[239] N. Waldron, D.-H. Kim, and J. del Alamo, "A self-aligned InGaAs HEMT architecture for logic applications," *IEEE Transactions on Electron Devices*, vol. 57, no. 1, pp. 297–304, 2010.

[240] X. Li, R. J. W. Hill, P. Longo, M. C. Holland, H. Zhou, S. Thoms, D. S. Macintyre, and I. Thayne, "Fully self-aligned process for fabricating 100 nm gate length enhancement mode GaAs metal-oxide-semiconductor field-effect transistors," *Journal of Vacuum Science Technology B: Microelectronics and Nanometer Structures*, vol. 27, no. 6, pp. 3153–3157, 2009.

[241] Z. Mahmood and S. Bashar, "Etching of p-type metal contact deposited on GaAs," in *International Conference on Transparent Optical Networks Proceedings*, 2001, pp. 317–320.

[242] O. Ignatova, S. Thoms, W. Jansen, D. Macintyre, and I. Thayne, "Lithography scaling issues associated with III-V MOSFETs," *Microelectronic Engineering*, vol. 87, no. 58, pp. 1049–1051, 2010.

[243] D. Tennant and A. Bleier, "Electron beam lithography of nanostructures," in *Handbook of Nanofabrication*, G. P. Wiederrecht, Ed. Elsevier, 2010, pp. 127–130.

[244] S. Thoms and D. S. Macintyre, "Long nanoscale gaps on III-V substrates by electron beam lithography," *Journal of Vacuum Science Technology B: Microelectronics and Nanometer Structures*, vol. 30, no. 6, p. 06F305, 2012.

[245] S. Thoms, "Electron beam lithography," in *Nanofabrication Handbook*, S. Cabrini and S. Kawata, Eds. CRC Press, Taylor and Francis Group, 2012, pp. 9–40.

[246] OpenSource, "ImageJ," 2013. [Online]. Available: http://rsb.info.nih.gov/ij/

[247] J. Malin, J. Beall, D. Zych, J. Bonkowski, S. Hillyard, N. Brette, K. Decker, and T. Nagle, "MESFET recess geometry impact on I_{dss}," in *International Conference on GaAs Manufacturing Technology* , 1998, pp. 83–86.

[248] M. Hagio, "Electrode reaction of GaAs metal semiconductor field-effect transistors in deionized water," *Journal of the Electrochemical Society*, vol. 140, no. 8, pp. 2402–2405, 1993.

[249] K. Cheng, "Effects of ohmic metal on electrochemical etching of GaAs in *p*HEMT manufacturing," in *International Conference on Compound Semiconductor Manufacturing Technology (CS MANTECH)*, 2009, pp. 13.3.1 –13.3.5.

[250] C. Kim, W. Lee, and K. Chung, "Threshold voltage control of the GaAs MESFET by varying dummy gate opening during recess etching," in *International Conference on GaAs Manufacturing Technology*, 1997, pp. 10–13.

[251] J. Crites, J. Dilley, S. Kittinger, W. Pickens, W. Polhamus, M. Balzan, and M. Drinkwine, "Electrochemical etching of ion implanted channel regions in GaAs," in *International Conference on Compound Semiconductor Manufacturing Technology (CS MANTECH)*, 2010, pp. 171–174.

[252] Z. Y. Liu, B. Hawkins, and T. F. Kuech, "Chemical and structural characterization of GaSb (100) surfaces treated by HCl-based solutions and annealed in vacuum," *Journal of Vacuum Science & Technology B*, vol. 21, no. 1, pp. 71–77, 2003.

[253] L. B. Ruppalt, E. R. Cleveland, J. G. Champlain, S. M. Prokes, J. B. Boos, D. Park, and B. R. Bennett, "Atomic layer deposition of Al_2O_3 on GaSb using in situ hydrogen plasma exposure," *Applied Physics Letters*, vol. 101, no. 23, p. 231601, 2012.

[254] K. Suzuki, Y. Harada, F. Maeda, K. Onomitsu, T. Yamaguchi, and K. Muraki, "Gate operation of InAs/AlGaSb heterostructures with an atomic-layer-deposited insulating layer," *Applied Physics Express*, vol. 4, no. 12, p. 125702, 2011.

[255] B. Brennan, M. Milojevic, C. Hinkle, F. Aguirre-Tostado, G. Hughes, and R. Wallace, "Optimisation of the ammonium sulphide $(NH_4)_2S$ passivation process on $In_{0.53}Ga_{0.47}As$," *Applied Surface Science*, vol. 257, no. 9, pp. 4082–4090, 2011.

[256] H. D. Trinh, E. Y. Chang, P. W. Wu, Y. Y. Wong, C. T. Chang, Y. F. Hsieh, C. C. Yu, H. Q. Nguyen, Y. C. Lin, K. L. Lin, and M. K. Hudait, "The influences of surface treatment and gas annealing conditions on the inversion behaviors of the atomic-layer-deposition Al_2O_3/n-$In_{0.53}Ga_{0.47}As$ metal-oxide-semiconductor capacitor," *Applied Physics Letters*, vol. 97, no. 4, p. 042903, 2010.

[257] I. Geppert, M. Eizenberg, A. Ali, and S. Datta, "Band offsets determination and interfacial chemical properties of the Al_2O_3/GaSb system," *Applied Physics Letters*, vol. 97, no. 16, p. 162109, 2010.

[258] S. McDonnell, D. Zhernokletov, A. Kirk, J. Kim, and R. Wallace, "In situ X-ray photoelectron spectroscopy characterization of Al_2O_3/GaSb interface evolution," *Applied Surface Science*, vol. 257, no. 20, pp. 8747–8751, 2011.

[259] A. Ali, H. S. Madan, A. P. Kirk, D. A. Zhao, D. A. Mourey, M. K. Hudait, R. M. Wallace, T. N. Jackson, B. R. Bennett, J. B. Boos, and S. Datta, "Fermi level unpinning of GaSb (100) using plasma enhanced atomic layer deposition of Al_2O_3," *Applied Physics Letters*, vol. 97, no. 14, p. 143502, 2010.

[260] A. Ali, H. Madan, A. P. Kirk, R. Wallace, D. Zhao, D. Mourey, M. Hudait, T. Jackson, B. Bennett, J. Boos, and S. Datta, "Fermi level unpinning of GaSb (100) using plasma enhanced ALD Al_2O_3 dielectric," in *Device Research Conference (DRC)*, 2010, pp. 27–30.

[261] Z. Tan, L. Zhao, J. Wang, and J. Xu, "Improved properties of HfO_2/Al_2O_3/GaSb MOS capacitors passivated with neutralized $(NH_4)_2S$ solutions," *ECS Solid State Letters*, vol. 2, no. 8, pp. P61–P62, 2013.

[262] L. Zhao, Z. Tan, R. Bai, N. Cui, J. Wang, and J. Xu, "Effects of sulfur passivation on GaSb metaloxidesemiconductor capacitors with neutralized and unneutralized $(NH_4)_2S$ solutions of varied concentrations," *Applied Physics Express*, vol. 6, no. 5, p. 056502, 2013.

[263] S. Gu, J. Min, Y. Taur, and P. M. Asbeck, "Characterization of interface defects in ALD Al_2O_3/p-GaSb MOS capacitors using admittance measurements in range from kHz to GHz," *Solid-State Electronics*, vol. 118, pp. 18–25, 2016.

[264] D. M. Zhernokletov, H. Dong, B. Brennan, J. Kim, and R. M. Wallace, "Optimization of the ammonium sulfide $(NH_4)_2S$ passivation process on InSb (111)A," *Journal of Vacuum Science & Technology B: Microelectronics and Nanometer Structures*, vol. 30, no. 4, p. 04E103, 2012.

[265] E. Hwang, C. Eaton, S. Mujumdar, H. Madan, A. Ali, D. Bhatia, S. Datta, and J. Ruzyllo, "Processing and characterization of GaSb/high-k dielectric interfaces," *ECS Transactions*, vol. 41, no. 5, pp. 157–162, 2011.

[266] C. Wang, M. Xu, J. Gu, D. W. Zhang, and P. D. Ye, "GaSb metal-oxide-semiconductor capacitors with atomic-layer-deposited HfAlO as gate dielectric," *Electrochemical and Solid-State Letters*, vol. 15, no. 3, pp. H51–H54, 2012.

[267] A. Greene, S. Madisetti, P. Nagaiah, M. Yakimov, V. Tokranov, R. Moore, and S. Oktyabrsky, "Improvement of the GaSb/Al_2O_3 interface using a thin InAs surface layer," *Solid-State Electronics*, vol. 78, pp. 56–61, 2012.

[268] L. B. Ruppalt, E. R. Cleveland, J. G. Champlain, B. R. Bennett, J. B. Boos, and S. M. Prokes, "Electronic properties of atomic-layer-deposited high-k dielectrics on GaSb(001) with hydrogen plasma pretreatment," *Journal of Vacuum Science & Technology B*, vol. 33, no. 4, p. 04E102, 2015.

[269] L. B. Ruppalt, E. R. Cleveland, J. G. Champlain, B. R. Bennett, and S. M. Prokes, "Integration of atomic layer deposited high-k dielectrics on GaSb via hydrogen plasma exposure," *AIP Advances*, vol. 4, no. 12, p. 127153, 2014.

[270] C. Merckling, X. Sun, A. Alian, G. Brammertz, V. V. Afanas'ev, T. Y. Hoffmann, M. Heyns, M. Caymax, and J. Dekoster, "GaSb molecular beam epitaxial growth on p-InP (001) and passivation with in situ deposited Al_2O_3 gate oxide," *Journal of Applied Physics*, vol. 109, no. 7, p. 073719, 2011.

[271] R.-L. Chu, W.-J. Hsueh, T.-H. Chiang, W.-C. Lee, H.-Y. Lin, T.-D. Lin, G. J. Brown, J.-I. Chyi, T.-S. Huang, T.-W. Pi, J. R. Kwo, and M. Hong, "Surface passivation of GaSb(100) using molecular beam epitaxy of Y_2O_3 and atomic layer deposition of Al_2O_3: a comparative study," *Applied Physics Express*, vol. 6, no. 12, p. 121201, 2013.

[272] T. Gotow, S. Fujikawa, H. I. Fujishiro, M. Ogura, T. Yasuda, and T. Maeda, "Effects of HCl treatment and predeposition vacuum annealing on Al_2O_3/GaSb/GaAs metaloxidesemiconductor structures," *Japanese Journal of Applied Physics*, vol. 54, no. 2, p. 021201, 2015.

[273] M. Yokoyama, Y. Asakura, H. Yokoyama, M. Takenaka, and S. Takagi, "Impact of process temperature on GaSb metal-oxide-semiconductor interface properties fabricated by ex-situ process," *Applied Physics Letters*, vol. 104, no. 26, p. 262901, 2014.

[274] M. Yokoyama, Y. Asakura, H. Yokoyama M. Takenaka, and S. Takagi, "Impact of Al_2O_3 ALD temperature on Al_2O_3/GaSb metal-oxide-semiconductor interface properties," in *Indium Phosphide and Related Materials (IPRM)*, 2013, pp. 1–2.

[275] B. Rajamohanan, D. Mohata, Y. Zhu, M. Hudait, Z. Jiang, M. Hollander, G. Klimeck, and S. Datta, "Design, fabrication, and analysis of p-channel arsenide/antimonide heterojunction tunnel transistors," *Journal of Applied Physics*, vol. 115, no. 4, p. 044502, 2014.

[276] Z. Yuan, C. Y. Chen, A. Kumar, A. Nainani, B. R. Bennett, J. B. Boos, and K. C. Saraswat, "Effects of oxidant dosing on GaSb (100) prior to atomic layer deposition and high-performance antimonide-based P-channel MOSFETs with Ni-alloy S/D," in *Device Research Conference (DRC)*, 2013, pp. 25–26.

[277] V. Tokranov, S. Madisetti, M. Yakimov, P. Nagaiah, N. Faleev, and S. Oktyabrsky, "All in-situ GaSb MOS structures on GaAs (001): growth, passivation and high-k oxides," *Journal of Crystal Growth*, vol. 378, pp. 631–635, 2013.

[278] J. A. Robinson and S. E. Mohney, "Characterization of sulfur passivated n-GaSb using transmission electron microscopy and the influence of passivation on ohmic contact resistance," *Journal of Applied Physics*, vol. 96, no. 5, pp. 2684–2688, 2004.

[279] E. Papis, A. Kudla, T. Piotrowski, K. Golaszewska, E. Kaminska, and A. Piotrowska, "Ellipsometric investigations of (100) GaSb surface under chemical etching and sulfide treatment," *Materials Science in Semiconductor Processing*, vol. 4, no. 13, pp. 293–295, 2001.

[280] G. Snider, 2013. [Online]. Available: https://www3.nd.edu/~gsnider/

[281] H.-D. Trinh, G. Brammertz, E. Chang, C. Kuo, C.-Y. Lu, Y. Lin, H. Q. Nguyen, Y. Y. Wong, B. Tran, K. Kakushima, and H. Iwai, "Electrical characterization of Al_2O_3/n-InAs metal-oxide-semiconductor capacitors with various surface treatments," *IEEE Electron Device Letters*, vol. 32, no. 6, pp. 752–754, 2011.

[282] H. Trinh, Y. Lin, E. Chang, C.-T. Lee, S.-Y. Wang, H. Nguyen, Y. Chiu, Q. Luc, H.-C. Chang, C.-H. Lin, S. Jang, and C. Diaz, "Electrical characteristics of Al_2O_3/InSb MOSCAPs and the effect of postdeposition annealing temperatures," *IEEE Transactions on Electron Devices*, vol. 60, no. 5, pp. 1555–1560, 2013.

[283] S. Svensson, D. Donetsky, D. Wang, H. Hier, F. Crowne, and G. Belenky, "Growth of type II strained layer superlattice, bulk InAs and GaSb materials for minority lifetime characterization," *Journal of Crystal Growth*, vol. 334, no. 1, pp. 103–107, 2011.

[284] D. Martn and C. Algora, "Temperature-dependent GaSb material parameters for reliable thermophotovoltaic cell modelling," *Semiconductor Science and Technology*, vol. 19, no. 8, p. 1040, 2004.

[285] C. L. Hinkle, A. M. Sonnet, E. M. Vogel, S. McDonnell, G. J. Hughes, M. Milojevic, B. Lee, F. S. Aguirre-Tostado, K. J. Choi, H. C. Kim, J. Kim, and R. M. Wallace, "GaAs interfacial self-cleaning by atomic layer deposition," *Applied Physics Letters*, vol. 92, no. 7, p. 071901, 2008.

[286] C. H. Hou, M. C. Chen, C. H. Chang, T. B. Wu, and C. D. Chiang, "Interfacial cleaning effects in passivating InSb with Al_2O_3 by atomic layer deposition," *Electrochemical and Solid-State Letters*, vol. 11, no. 6, pp. D60–D63, 2008.

[287] ITRS, "Process integration, devices, and structures," 2011. [Online]. Available: http://www.itrs.net/Links/2011ITRS/Home2011.htm

[288] C. L. Hinkle, A. M. Sonnet, M. Milojevic, F. S. Aguirre-Tostado, H. C. Kim, J. Kim, R. M. Wallace, and E. M. Vogel, "Comparison of n-type and p-type GaAs oxide growth and its effects on frequency dispersion characteristics," *Applied Physics Letters*, vol. 93, no. 11, p. 113506, 2008.

[289] L. Morassi, A. Padovani, G. Verzellesi, D. Veksler, I. Ok, and G. Bersuker, "Interface-trap effects in inversion-type enhancement-mode InGaAs/ZrO_2 n-channel MOSFETs," *IEEE Transactions on Electron Devices*, vol. 58, no. 1, pp. 107–114, 2011.

[290] B. Benbakhti, J. Ayubi-Moak, K. Kalna, D. Lin, G. Hellings, G. Brammertz, K. D. Meyer, I. Thayne, and A. Asenov, "Impact of interface state trap density on the performance characteristics of different III-V MOSFET architectures," *Microelectronics Reliability*, vol. 50, no. 3, pp. 360–364, 2010.

[291] N. V. Nguyen, M. Xu, O. A. Kirillov, P. D. Ye, C. Wang, K. Cheung, and J. S. Suehle, "Band offsets of $Al_2O_3/In_xGa_{1-x}As$ (x=0.53 and 0.75) and the effects of postdeposition annealing," *Applied Physics Letters*, vol. 96, no. 5, p. 052107, 2010.

[292] V. V. Afanas'ev, A. Stesmans, G. Brammertz, A. Delabie, S. Sionke, A. O'Mahony, I. M. Povey, M. E. Pemble, E. O'Connor, P. K. Hurley, and S. B. Newcomb, "Energy barriers at interfaces between (100) $In_xGa_{1-x}As$ ($0 \leq x \leq 0.53$) and atomic-layer deposited Al_2O_3 and HfO_2," *Applied Physics Letters*, vol. 94, no. 20, 2009.

[293] J. Ahn, B. Shin, and P. C. McIntyre, "In-situ As_2 decapping and atomic layer deposition of Al_2O_3 on n-InGaAs (100)," *ECS Transactions*, vol. 45, no. 4, pp. 183–188, 2012.

[294] B. Shin, J. B. Clemens, M. A. Kelly, A. C. Kummel, and P. C. McIntyre, "Arsenic decapping and half cycle reactions during atomic layer deposition of Al_2O_3 on $In_{0.53}Ga_{0.47}As(001)$," *Applied Physics Letters*, vol. 96, no. 25, p. 252907, 2010.

[295] R. Chau, S. Datta, M. Doczy, B. Doyle, B. Jin, J. Kavalieros, A. Majumdar, M. Metz, and M. Radosavljevic, "Benchmarking nanotechnology for high-performance and low-power logic transistor applications," *IEEE Transactions on Nanotechnology*, vol. 4, no. 2, pp. 153–158, 2005.

[296] J. Lee, W. Tsai, F. Zhu, S. Koveshnikov, H. Bakhru, R. Kambhampati, V. Tokranov, M. Yakimov, and S. Oktyabrsky, "Challenges and progress in III-V MOSFETs for CMOS circuits," *International Journal of High Speed Electronics and Systems*, vol. 18, no. 04, pp. 761–772, 2008.

[297] V. Djara, T. O'Regan, K. Cherkaoui, M. Schmidt, S. Monaghan, E. O'Connor, I. Povey, D. O'Connell, M. Pemble, and P. Hurley, "Electrically active interface defects in the $In_{0.53}Ga_{0.47}As$ MOS system," *Microelectronic Engineering*, vol. 109, pp. 182–188, 2013.

[298] G. Brammertz, A. Alian, D.-C. Lin, M. Meuris, M. Caymax, and W. E. Wang, "A combined interface and border trap model for high-mobility substrate metal-oxide-semiconductor devices applied to $In_{0.53}Ga_{0.47}As$ and InP capacitors," *IEEE Transactions on Electron Devices*, vol. 58, no. 11, pp. 3890–3897, 2011.

[299] B. Shin, J. R. Weber, R. D. Long, P. K. Hurley, C. G. V. de Walle, and P. C. McIntyre, "Origin and passivation of fixed charge in atomic layer deposited aluminum oxide gate insulators on chemically treated InGaAs substrates," *Applied Physics Letters*, vol. 96, no. 15, p. 152908, 2010.

[300] J. Hu and H.-S. Philip Wong, "Effect of annealing ambient and temperature on the electrical characteristics of atomic layer deposition $Al_2O_3/In_{0.53}Ga_{0.47}As$ metal-oxide-semiconductor capacitors and MOSFETs," *Journal of Applied Physics*, vol. 111, no. 4, p. 044105, 2012.

[301] K. Roy, S. Mukhopadhyay, and H. Mahmoodi-Meimand, "Leakage current mechanisms and leakage reduction techniques in deep-submicrometer CMOS circuits," *Proceedings of the IEEE*, vol. 91, no. 2, pp. 305–327, 2003.

[302] D. Bol, R. Ambroise, D. Flandre, and J. D. Legat, "Interests and limitations of technology scaling for subthreshold logic," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 17, no. 10, pp. 1508–1519, 2009.

[303] A. O'Mahony, S. Monaghan, G. Provenzano, I. M. Povey, M. G. Nolan, E. O'Connor, K. Cherkaoui, S. B. Newcomb, F. Crupi, P. K. Hurley, and M. E. Pemble, "Structural and electrical analysis of the atomic layer deposition of HfO_2/n-$In_{0.53}Ga_{0.47}As$ capacitors with and without an Al_2O_3 interface control layer," *Applied Physics Letters*, vol. 97, no. 5, p. 052904, 2010.

[304] Y. Wang, Y.-T. Chen, H. Zhao, F. Xue, F. Zhou, and J. C. Lee, "Improved electrical properties of HfO_2-based gate dielectrics on InP substrate using Al_2O_3/HfO_2 and SF6 plasma treatment," *Electrochemical and Solid-State Letters*, vol. 14, no. 7, pp. H291–H293, 2011.

[305] L. K. Chu, C. Merckling, A. Alian, J. Dekoster, J. Kwo, M. Hong, M. Caymax, and M. Heyns, "Low interfacial trap density and sub-nm equivalent oxide thickness in $In_{0.53}Ga_{0.47}As$ (001) metal-oxide-semiconductor devices using molecular beam deposited HfO_2/Al_2O_3 as gate dielectrics," *Applied Physics Letters*, vol. 99, no. 4, 2011.

[306] Y. Xuan, Y. Wu, T. Shen, T. Yang, and P. Ye, "High performance submicron inversion-type enhancement-mode InGaAs MOSFETs with ALD Al_2O_3, HfO_2 and HfAlO as gate dielectrics," in *IEEE International Electron Devices Meeting (IEDM) Technical Digest*, 2007, pp. 637–640.

[307] S. Verdonckt-Vandebroek, E. Crabbe, B. Meyerson, D. Harame, P. Restle, J. M. C. Stork, and J. Johnson, "SiGe-channel heterojunction *p*-MOSFET's," *IEEE Transactions on Electron Devices*, vol. 41, no. 1, pp. 90–101, 1994.

[308] S. Mathew, G. Niu, W. Dubbelday, J. Cressler, J. Ott, J. Chu, P. Mooney, K. L. Kavanagh, B. Meyerson, and I. Lagnado, "Hole confinement and low-frequency noise in SiGe *p*FETs on silicon-on-sapphire," *IEEE Electron Device Letters*, vol. 20, no. 4, pp. 173–175, 1999.

[309] K. Michelakis, A. Vilches, C. Papavassiliou, S. Despotopoulos, K. Fobelets, and C. Toumazou, "Average drift mobility and apparent sheet-electron density profiles in strained-Si-SiGe buried-channel depletion-mode *n*-MOSFETs," *IEEE Transactions on Electron Devices*, vol. 51, no. 8, pp. 1309–1314, 2004.

[310] Z. Yuan, A. Nainani, B. R. Bennett, J. Brad Boos, M. G. Ancona, and K. C. Saraswat, "Amelioration of interface state response using band engineering in III-V quantum well metal-oxide-semiconductor field-effect transistors," *Applied Physics Letters*, vol. 100, no. 14, p. 143503, 2012.

[311] A. Ali, S. Mookerjea, E. Hwang, S. Koveshnikov, S. Oktyabrsky, V. Tokranov, M. Yakimov, R. Kambhampati, W. Tsai, and S. Datta, "HfO_2 gated, self aligned and directly contacted indium arsenide quantum-well transistors for logic applications - a temperature and bias dependent study," in *Device Research Conference (DRC)*, 2009, pp. 55–56.

[312] Y. Urabe, T. Yasuda, H. Ishii, T. Itatani, N. Miyata, H. Yamada, N. Fukuhara, M. Hata, M. Takenaka, and S. Takagi, "On the mechanisms limiting mobility in InP/InGaAs buried channel *n*MISFETs," *Microelectronic Engineering*, vol. 88, no. 7, pp. 1076–1078, 2011.

[313] N. Onojima, A. Kasamatsu, N. Hirose, T. Mimura, and T. Matsui, "C-V characterization of Schottky- and MIS-gate SiGe/Si HEMT structures," *Applied Surface Science*, vol. 254, no. 19, pp. 6162–6164, 2008.

[314] K. Martens, "Electrical characterisation and modelling of Ge/III-V - dielectric interfaces," Ph.D. dissertation, 2009.

[315] I. Krylov, A. Gavrilov, M. Eizenberg, and D. Ritter, "Correlation between Ga-O signature and midgap states at the $Al_2O_3/In_{0.53}Ga_{0.47}As$ interface," *Applied Physics Letters*, vol. 101, no. 6, p. 063504, 2012.

[316] S. Takagi, A. Toriumi, M. Iwase, and H. Tango, "On the universality of inversion layer mobility in Si MOSFET's: part I - effects of substrate impurity concentration," *IEEE Transactions on Electron Devices*, vol. 41, no. 12, pp. 2357–2362, 1994.

[317] Y. Xuan, Y. Wu, H. Lin, T. Shen, and P. Ye, "Submicrometer inversion-type enhancement-mode InGaAs MOSFET with atomic-layer-deposited Al_2O_3 as gate dielectric," *IEEE Electron Device Letters*, vol. 28, no. 11, pp. 935–938, 2007.

[318] M. A. Negara, D. Veksler, J. Huang, G. Ghibaudo, P. K. Hurley, G. Bersuker, N. Goel, and P. Kirsch, "Analysis of effective mobility and hall effect mobility in high-k based $In_{0.75}Ga_{0.25}As$ metal-oxide-semiconductor high-electron-mobility transistors," *Applied Physics Letters*, vol. 99, no. 23, p. 232101, 2011.

[319] C. Hinkle, A. Sonnet, R. Chapman, and E. Vogel, "Extraction of the effective mobility of $In_{0.53}Ga_{0.47}As$ MOSFETs," *Electron Device Letters, IEEE*, vol. 30, no. 4, pp. 316–318, 2009.

[320] W. Lu, X.-W. Wang, R. Hammond, A. Kuliev, S. Koester, J. Chu, K. Ismail, T.-P. Ma, and I. Adesida, "p-type SiGe transistors with low gate leakage using SiN gate dielectric," *IEEE Electron Device Letters*, vol. 20, no. 10, pp. 514–516, 1999.

[321] T. Irisawa, S. Koh, K. Nakagawa, and Y. Shiraki, "Growth of SiGe/Ge/SiGe heterostructures with ultrahigh hole mobility and their device application," *Journal of Crystal Growth*, vol. 251, no. 1-4, pp. 670–675, 2003.

[322] N. Taoka, M. Yokoyama, S. Kim, R. Suzuki, R. Iida, S. Lee, T. Hoshii, W. Jevasuwan, T. Maeda, T. Yasuda, O. Ichikawa, N. Fukuhara, M. Hata, M. Takenaka, and S. Takagi, "Impact of Fermi level pinning inside conduction band on electron mobility of $In_xGa_{1-x}As$ MOSFETs and mobility enhancement by pinning modulation," in *IEEE International Electron Devices Meeting (IEDM) Technical Digest*, 2011, pp. 610–613.

[323] N. Taoka, M. Yokoyama, S. Hyeon Kim, R. Suzuki, S. Lee, R. Iida, T. Hoshii, W. Jevasuwan, T. Maeda, T. Yasuda, O. Ichikawa, N. Fukuhara, M. Hata, M. Takenaka, and S. Takagi, "Impact of Fermi level pinning inside conduction band on electron mobility in InGaAs metal-oxide-semiconductor field-effect transistors," *Applied Physics Letters*, vol. 103, no. 14, p. 143509, 2013.