DESIGN AND SIMULATION OF
PLANAR ELECTRONIC NANODEVICES
FOR TERAHERTZ AND MEMORY
APPLICATIONS

A thesis submitted to The University of Manchester for the degree of
Doctor of Philosophy
in the Faculty of Engineering and Physical Sciences

2013

MUBARAK ALI
SCHOOL OF ELECTRICAL AND ELECTRONIC ENGINEERING
To my Parents Zahabar Ali & Ramzan Beevi,
Wife Waheeda
& the Jewel of our lives Isra Laieqa.
Contents

Abstract vi
Declaration vii
Copyright vii
Acknowledgements viii
List of Tables ix
List of Figures x
List of Abbreviations xix

1 Introduction
1.1 The Terahertz gap .................................................................3
1.2 Semiconductor memories .....................................................6
1.3 Computational electronics ..................................................8
1.4 Motivation and aims .........................................................11

2 Background
2.1 Terahertz (THz) technology ................................................13
2.1.1 THz applications .......................................................14
2.1.2 Time-domain spectroscopy ...........................................20
2.2 THz generation .................................................................21
2.2.1 Optical approaches ......................................................22
2.2.2 Electronic approaches ................................................25
2.2.3 Solid-state devices .......................................................26
2.3 The Gunn effect ...............................................................32
2.3.1 Physics ..................................................................32
2.3.2 Conventional Gunn diode and limitations ......................40
## Contents

2.3.3 Planar Gunn devices .................................................. 41
2.4 Self-switching device (SSD) ..................................................... 43
2.4.1 Applications ................................................................ 45
2.4.2 Gunn oscillations ........................................................ 47
2.4.3 Memory effect............................................................. 49

3 Development of ATLAS for planar nanodevices

3.1 Introduction .............................................................................. 52
3.2 SILVACO ................................................................................. 53
3.3 Device physics ......................................................................... 56
3.3.1 Poisson's equation ...................................................... 57
3.3.2 Carrier continuity equations ........................................ 57
3.3.3 Transport equations .................................................... 58
3.4 Physical models ....................................................................... 58
3.4.1 Carrier statistics and transport .................................... 59
3.4.2 Mobility ....................................................................... 63
3.4.3 Carrier generation-recombination models................... 65
3.5 Model validation ....................................................................... 67

4 Planar Gunn diodes

4.1 Introduction .............................................................................. 69
4.2 DC Conditions ........................................................................ 69
4.3 Electron distribution in the device ............................................ 71
4.4 Gunn oscillations ..................................................................... 76
4.4.1 Domain profiles........................................................... 76
4.4.2 Threshold bias ............................................................ 79
4.4.3 Cessation bias ............................................................ 80
4.4.4 Factors affecting oscillating performance ................... 81
4.5 Discussion and conclusions ..................................................... 93
## 5 Planar nano-memory devices

5.1 Introduction ..............................................................................94
5.2 Conventional memory devices .................................................95
  5.2.1 Flash memory .......................................................................96
  5.2.2 Dynamic Random Access Memory (DRAM) .........................100
  5.2.3 New developments ......................................................102
5.3 Planar nano-memory devices ................................................104
  5.3.1 Surface states and space charge ...................................108
  5.3.2 Current transport in Schottky barriers .........................111
5.4 Simulation results ....................................................................115
  5.4.1 Material, models and parameters ..................................116
  5.4.2 Steady-state characteristics ...........................................117
  5.4.3 Transient characteristics ...............................................118
  5.4.4 Array ..............................................................................133
  5.4.5 Electrical characteristics ..............................................135
  5.4.6 Oxide-filled trenches ....................................................135
5.5 Discussion and conclusions ...................................................136

## 6 Summary and outlook

6.1 Summary................................................................................138
6.2 Outlook...................................................................................141

**Bibliography**

**Appendix A**  
Gunn Diode Simulation Code

**Appendix B**  
pRAM Simulation Code
Abstract

The performances derived from current electronic technology are fast approaching a plateau since traditional vertically-layered devices are already in the scaling-limit range. The prospects of using planar devices as a solution have become increasingly promising. Besides, they provide additional advantages of being simple yet operating at very high speeds. In this study, the feasibility of utilising a planar nanoscale unipolar diode or a self-switching device (SSD) for terahertz emission and memory applications is demonstrated using simulations. Detailed characterisation of the devices is performed, paying close attention to their geometrical parameters and the surface-charge density which are crucial in planar electron transport.

The emission from the SSD is profiled using electron dynamics in the device evidencing the presence of Gunn domains that lead to current oscillations. Following this, the device performance as determined by lithography-tuneable parameters of channel-length, channel-width and interface-charge density is investigated, in terms of their oscillating frequencies and current amplitudes. The study shows that the geometrical dimensions of the SSD can be tailored for optimum emission frequency and current oscillation magnitude, simply by altering the length and width of the channel, respectively. The highest fundamental frequency attained is 0.2 THz and higher harmonics could achieve up to 1 THz. Moreover, the interface-charge density has a much greater effect on the oscillation frequency than expected, providing some promise to extend emission frequency to a range that has been difficult to achieve using a solid-state device at room temperature.

The flexibility of the SSD design has been further exploited to conceptualise a novel planar memory device aimed at overcoming the stagnated processing speeds of multilayer computational chips relying on interconnects. The structure is based on a high surface-to-volume ratio which enables conduction to be controlled by a memory storage region that can be charged and discharged by a control gate. Initially, structure dimensions are tuned and thereafter, the memory retention times resulting from the optimisation of geometrical and electrical parameters are discussed. The energy consumption of the device is much lower than flash memories, potentially useful for emerging low-power applications, particularly when device arrays are designed.
Declaration

No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright

1. The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the “Copyright”) and s/he has given The University of Manchester certain rights to use such Copyright including for administrative purposes.

2. Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made only in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.

3. The ownership of certain Copyright, patents, designs, trademarks and other intellectual property (the “Intellectual Property”) and any reproductions of copyright works in the thesis, for example graphs and tables (“Reproductions”), which may be described in this thesis, may not be owned by the author and may be owned by the third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

4. Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in The University IP Policy:

http://documents.manchester.ac.uk/DocuInfo.aspx?DocID=487,

The University Library’s regulations:

http://www.manchester.ac.uk/library/aboutus/regulations) and in

The University’s policy on the Presentation of Theses.
Acknowledgements

All praise is due to Allah, whose mercy and guidance was of utmost importance to the completion of this thesis.

First and foremost, I wish to express my humblest gratitude to my parents who till today have not taken their hands off the love paddle in nurturing and encouraging the three of us to pursue our dreams and ambitions.

A salute of admiration goes out to all educators who have given selflessly toward the cause of moulding the future and in the process helping to shape mine. I am especially grateful for the supervision provided by Professor Aimin Song, for his invaluable guidance, support, time and encouragement during this project.

Colleagues who have become friends: you have painted life’s dull corners in a myriad of colours. Thanks for the enjoyable discussions and post-work banters.

This work would also not have been possible without the support of the ‘Science 2015’ Alumni Fund scholarship which I am truly grateful to.

Last but not least, an emphatic appreciation to my family who have been a pillar of support, especially my wife, Waheeda and daughter, Isra Laieqa, the strength and joy in this adventure.
List of Tables

3.1 Summary of the physical models used in the Gunn device simulations...... 67

4.1 Working bias range for different channel lengths................................. 81
4.2 Working bias range for different channel widths................................. 81

5.1 On/Off ratios and memory retention times for gap dimensions of a 1:1 ratio ranging from 20 to 100 nm. ................................................................. 121
5.2 On/Off ratios and memory retention times for combinations of gap sizes. 124
5.3 Ranking of memory retention times (a) $t_1$ and (b) $t_2$............................. 125
5.4 Average barrier heights per nanometer of combined gap size relating to the dependency of $t_1$. ................................................................. 125
5.5 Pulse width dependence of On/Off ratios and memory retention times.

\[ |V_{G2}| = 10V. \] ................................................................................. 127
List of Figures

1.1 Intel CPU trends. Moore’s Law is expected to continue for at least a few more years, but raw clock cycles have met their limit. Any future performance gains are going to be accomplished in fundamentally different ways for at least the next couple of processor generations. Most current applications will no longer benefit without significant redesign [4].

1.2 Representation of the electromagnetic spectrum illustrating the THz gap relative to the microwave and infrared.

1.3 Global market for THz radiation devices and systems through 2021, (a) by type of system and (b) by application. Source: BCC Research.

1.4 Sales figures showing the growth in the flash memory unit shipments overtaking that of DRAM. Source: Market research firm, IC Insights Inc.

2.1 The electromagnetic spectrum depicting the THz region.

2.2 The Planck instrument (left) and THz observation of the cosmic microwave background (right).

2.3 (a) Terahertz narcotic detection: THz image (top) and photograph (bottom) of specimens under inspection. The three kinds of powder hidden in the envelopes are distinguished using THz-TDS by their own THz signatures. Image courtesy of RIKEN, Japan. (b) ThruVision’s passive imaging security system detects terahertz radiation emitted from objects hidden on the body. Source: ThruVision Ltd.

2.4 TeraView’s Terahertz Pulsed Imaging (TPITM) is able to distinguish between the different types of tissue in a human tooth; detect caries at an early stage in the enamel layers of human teeth and monitor early erosion of the enamel at the surface of the tooth. (a) X-ray image. (b) THz image. Source: TeraView Ltd.

2.5 THz absorption spectra of barbital. (a) THz-TDS. (b) FT-FIR. The polymorph sample (form B) was obtained by heating original barbital (form A) at 160 °C for 30 minutes [65].

2.6 Time-domain spectroscopy (TDS) setup.

2.7 Generation of THz radiation using optical and electronic techniques.
### List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.8</td>
<td>Schematic of a heterodyne THz source. Top: Beams from two lasers are combined and the resultant THz beat focussed onto the antenna which radiates the THz waves through a hemispherical lens to better couple it to free space. Bottom: Spiral antenna on a photoconductive (PC) GaAs substrate. A bias is placed between the photomixer electrodes so that electrons and holes generated by the THz beat are attracted to opposite sides [82].</td>
</tr>
<tr>
<td>2.9</td>
<td>(a) Schematic of a semiconductor source of THz waves based on a QCL design.</td>
</tr>
<tr>
<td>2.10</td>
<td>Schematic of a free electron laser [95].</td>
</tr>
<tr>
<td>2.11</td>
<td>I-V of a device showing negative differential resistance [105].</td>
</tr>
<tr>
<td>2.12</td>
<td>State of the art results from GaAs and InP Gunn devices. Numbers next to the symbols denote dc to RF conversion efficiencies in percent [104].</td>
</tr>
<tr>
<td>2.13</td>
<td>(a) 3D view of a MMIC power amplifier in a package [120]. (b) Scanning Electron Microscope image of a planar GaAs-based MMIC working at 0.7THz to be integrated with amplifiers under the DARPA Terahertz Transistor and Imaging Program. This represents the first wafer-bonded all-planar mixers and amplifiers for a 0.7 THz receiver system [120].</td>
</tr>
<tr>
<td>2.14</td>
<td>The output power performance of leading electronic source technology. The multiplexer is a Schottky barrier diode frequency multiplier. The figure clearly shows the persisting limitations within the THz gap, and the power efficiencies (not shown) are also very low, which is a problem for many practical applications [53, 58].</td>
</tr>
<tr>
<td>2.15</td>
<td>Carrier drift velocity vs electric field for three semiconductor materials: GaAs, InP and GaN [121].</td>
</tr>
<tr>
<td>2.16</td>
<td>Electron occupations for (a) GaAs, (b) InP and (c) GaN [112].</td>
</tr>
<tr>
<td>2.17</td>
<td>Electron occupations under various electric field levels for n-GaAs [112].</td>
</tr>
<tr>
<td>2.18</td>
<td>Negative differential mobility region [112].</td>
</tr>
<tr>
<td>2.19</td>
<td>Stable dipole domain formation with the growth of space charge [122].</td>
</tr>
<tr>
<td>2.20</td>
<td>Conventional Gunn diode structure.</td>
</tr>
<tr>
<td>2.21</td>
<td>Schematic of electron flow in (a) a vertical Gunn diode and (b) a planar Gunn diode.</td>
</tr>
</tbody>
</table>
List of Figures

2.22 (a) A typical \textit{I-V} characteristic of the self-switching diode (SSD). The SSD was fabricated on InGaAs/InP substrate with a 1.2 μm channel length ($L$) and 80 nm channel width ($W$). The measurement was performed at a low temperature of 4.2 K. (b) Illustration of the depletion regions at the boundaries caused by the charges at the surface states of the etched trenches. Depending on the sign of the applied voltage $V$, the effective channel width will (c) reduce or (d) increase resulting in the diode-like characteristic [46]

2.23 \textit{I-V} characteristics of two SSDs with different $W$ and $L = 1.2$ μm. Both SSDs were fabricated on InGaAs/InAlAs substrate. Turn-on voltage decreases as $W$ increases from 60 nm to 70 nm [46]

2.24 (a) Atomic-force micrograph of two SSDs connected in parallel fabricated using electron beam lithography. The channel length and width are 1.5 μm and 130 nm, respectively[48]. (b) SEM image of 100 SSDs connected in parallel, fabricated using a single lithography step without interconnection layers [131]

2.25 (a) Schematic top view (not to scale) of the simulated SSD. (b) Current responses of a SSD compared to a symmetric structure. The applied voltage changes from 2.5 to 3.0 V at time zero [50]

2.26 (a) Geometry of the InGaAs SSD. (b) Current sequence for Δ$V = 0.5$ V applied every 50 ps [155]

2.27 Hysteresis effect in the \textit{I-V} characteristic of a typical self-switching diode (SSD) that can be utilised for memory operations. Both dashed and solid arrows indicate the sweeping direction of the applied voltage. The measurement was carried out at temperature of 24 K [52]

2.28 The design of a novel planar memory device

3.1 ATLAS inputs and outputs [172]

3.2 Schematic of the SSD device based on an In$_{0.53}$Ga$_{0.47}$As/In$_{0.53}$Al$_{0.47}$As heterostructure, where a 2DEG is formed at the heterointerface [158]

3.3 \textit{I-V} characteristics compared between ATLAS simulation and the Monte Carlo simulation [144, 145] of two SSDs with channel width $W = 50$nm having different channel lengths: (i) $L = 0.1$μm and (ii) $L = 1.0$ μm

3.4 Current response of a SSD with channel length $L=1.25$ μm and channel width $W=60$ nm using (a) ATLAS and (b) Monte Carlo simulations
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>Geometry of the simulated SSD in microns, not drawn to scale.</td>
<td>70</td>
</tr>
<tr>
<td>4.2</td>
<td>$I-V$ characteristics of an SSD with $L = 1.2 , \mu m$, $W = 80 , nm$ for different interface-charge densities.</td>
<td>71</td>
</tr>
<tr>
<td>4.3</td>
<td>Comparison of the experimental $I-V$ characteristic of a SSD with $L = 1.5 , \mu m$ and $W = 450 , nm$ with the simulated $I-V$ performed using ATLAS.</td>
<td>72</td>
</tr>
<tr>
<td>4.4</td>
<td>The electron concentration per $cm^3$ at $0V$ in the channel; altered by the introduction of interface-charge density.</td>
<td>72</td>
</tr>
<tr>
<td>4.5</td>
<td>(a) Contour plots and (b) profile lines of electron density for incremental positive and negative biases.</td>
<td>74</td>
</tr>
<tr>
<td>4.6</td>
<td>(a) Calculated $I-V$ characteristic of an InGaAs SSD with a $L = 1.0 , \mu m$ and $W = 70 , nm$ and (b) average electron density and velocity in the InGaAs SSD channel versus applied bias [154].</td>
<td>75</td>
</tr>
<tr>
<td>4.7</td>
<td>Average electron density versus applied bias simulated by ATLAS for the SSD in Fig. 4.1 ($L = 1.2 , \mu m$ and $W = 80 , nm$).</td>
<td>75</td>
</tr>
<tr>
<td>4.8</td>
<td>Current response from the SSD when biased in incremental bias steps for (a) $L = 1.2 , \mu m$, $W = 80 , nm$; and (b) $L = 1.2 , \mu m$, $W = 100 , nm$.</td>
<td>77</td>
</tr>
<tr>
<td>4.9</td>
<td>(a) Current oscillation from Fig. 4.8b at $V = 2.5 , V$ showing 6 time moments in one period.</td>
<td>78</td>
</tr>
<tr>
<td>4.10</td>
<td>(a) Electron density, (b) electric field, (c) electron energy and (d) electron velocity line profiles at the entrance of the channel showing the initiation of domains above the threshold voltage.</td>
<td>79</td>
</tr>
<tr>
<td>4.11</td>
<td>The current oscillations showing the threshold and cessation biases for (a) the reference geometry: $L = 1.2 , \mu m$, $W = 100 , nm$, (b) the longer channel: $L = 1.8 , \mu m$, $W = 100 , nm$; and (c) a wider channel $L = 1.2 , \mu m$, $W = 180 , nm$.</td>
<td>80</td>
</tr>
<tr>
<td>4.12</td>
<td>(a) Frequencies and amplitudes of the fundamental frequencies for increasing channel lengths, $W = 80 , nm$, $\sigma = 4.0\times10^{11} , cm^{-2}$ and $V = 2.5 , V$. The channel length has a stronger influence on the frequency than the amplitude of current oscillations. (b) The inverse relation of the fundamental frequency with the length of the channel.</td>
<td>82</td>
</tr>
<tr>
<td>4.13</td>
<td>Frequency spectra of the SSDs with increasing channel lengths from 1.0 to 2.5 $\mu m$; and their waveforms are shown as insets. A greater number of harmonics within a narrow frequency range become possible owing to the nature of the waveforms.</td>
<td>83</td>
</tr>
</tbody>
</table>
4.14 The current oscillation for a long channel ($L = 1.8 \, \mu m$) analysed in three sections by plotting their electron density distributions. The dip in section (ii) occurs owing to the increased domain propagation distance and the insufficient bias (2.5 V) to drive the domains efficiently through the middle section of the channel. ............................................................... 85

4.15 Frequencies and amplitudes of the fundamental frequencies for increasing channel widths, $L = 1.2 \, \mu m$ and $\sigma = 4.0 \times 10^{11} \, \text{cm}^{-2}$. The results are extracted from their respective threshold voltages which increase proportionately from 1.5 to 6.5 V. The channel width has a stronger influence on the amplitudes of oscillations than their frequencies. ............................................................... 85

4.16 Frequency spectra of the SSDs with increasing channel widths from 100 to 200 nm; and their waveforms are shown as insets. The magnitude of the fundamental harmonic peak represents the amplitude of oscillations. A greater number of harmonics within a broad frequency range is achievable in wider channels. ............................................................................... 86

4.17 (a) A greater number of electrons are contained in the domains of channels that are wider ($W = 180 \, \text{nm}$). (b) The electric field domains are correspondingly larger meaning a larger bias is required for them to traverse the channel. ......................................................................................... 87

4.18 Frequencies and amplitudes of the fundamental frequencies for increasing interface-charge densities, $L = 1.2 \, \mu m$ and $V = 2.5 \, \text{V}$. The widths of the channels are increased to accommodate similar average currents in the channels. A strong influence on the frequency of emission is observed. ..... 88

4.19 Frequency spectra of the SSDs with increasing interface-charge densities from $2 \times 10^{11}$ to $10 \times 10^{11} / \text{cm}^2$ and their waveforms are shown as insets. The high interface-charge density device produces the highest fundamental frequency and the fourth harmonic reaches 0.9 THz. ........................................ 89

4.20 The electron density domain initiation profiles show that the domain forms further inside the channel as the interface-charge number density is increased. The effective propagation distance is reduced and the frequency is enhanced ........................................................................................................ 90
List of Figures

4.21 (a) The waveforms of the devices showing their relative frequencies and amplitudes of oscillations with respect to geometry and their Fourier transformed emission peaks at (i) the threshold bias and (ii) the highest bias before cessation for (b) the reference geometry: L = 1.2 µm, W = 80 nm, (c) the longer channel: L = 1.8 µm, W = 100 nm; and (d) a wider channel L = 1.2 µm, W = 180 nm................................................................. 91

4.22 First, second and third harmonics data for the geometrical variations in (a) channel length, (b) channel width; and (c) interface-charge density. ........ 92

5.1 Schematic cross-section representation of an n-channel MOS transistor structure [19]. .......................................................................................... 96

5.2 A generic CMOS flash memory transistor [17].......................................................... 97

5.3 (a) Read and write/erase operation principles of a flash memory. The schematic of the memory transistor illustrates the bias conditions that are employed for the case of CHE injection or F–N substrate electron injection write modes, and for the case of source-side electron extraction erase mode. (b) Circuit schematic of a one-transistor flash memory cell in a memory configuration (two memory cells are shown) [17]........................................ 97

5.4 Schematic of (a) conventional planar one-transistor one capacitor DRAM cell, and (b) DRAM circuit in a memory-array configuration (two memory cells are shown) [6, 17]. .......................................................................................... 101

5.5 The structures of (a) a floating gate nonvolatile memory; and (b) a nanocrystal nonvolatile memory [218]. ...................................................... 103

5.6 Atomic force microscope image of novel SR-latch proposed by Sun et. al. [226]........................................................................................................ 105

5.7 (a) Low temperature I-V characteristics of a self-switching memory. (b) Illustration of surface state charging and discharging leading to SSD memory effect. When the reverse bias exceeds Vth−, the surface states discharge by charge transfer into the channel (reducing the surface depletion region). Conversely, when the forward bias exceeds Vth+, the surface states are recharged (increasing the surface depletion region) [52]. ................................................................. 106

5.8 The proposed memory device structure. The conduction band profile along the dotted line is illustrated in Fig. 5.9. .............................................. 106
List of Figures

5.9 The conduction band schematic in the pRAM along the dotted line shown in Fig. 5.8. Memory charge extends the depletion region in the channel inhibiting current....................................................................................... 108

5.10 Schematic electron energy levels near the surface of a clean semiconductor: (a) disequilibrium and (b) equilibrium [230]......................................................... 109

5.11 Schematic diagrams showing the energy levels and free charge carrier densities (logarithmic scale) from the n-type semiconductor surface to the bulk. The blue dotted lines indicate the corresponding space charge region of thickness, D. $n_e =$ free electron density; $n_h =$ free hole density; $n_i =$ intrinsic carrier density [227].................................................................................. 110

5.12 (a) GaAs substrate with 200 nm width insulating trenches (air) with a 100 nm gap between the trenches. (b) Conduction band diagram along the dotted line showing the magnitude of band bending for different interface-charge densities................................................................................................ 111

5.13 Energy band diagrams of metal and n-type semiconductor contacts. $E_{\text{vac}} =$ vacuum energy, $E_c =$ energy of conduction band minimum, $E_v =$ energy of valence band maximum, $\varphi_m =$ metal work function, $\varphi_s =$ semiconductor work function, $\chi_s =$ electron affinity of the semiconductor [217]. ............ 112

5.14 Current transport for a range of Schottky barrier heights using Eq. (5.6) for a contact area of depth = 200 nm and width = 50 nm of a Silicon substrate.114

5.15 (a) Conduction band contour plot for a Silicon substrate with insulating trenches creating a nanogap. (b) Conduction band diagram along the dotted line showing the magnitude of potential barriers for different gap sizes. Potential barriers of ~0.8 eV are attainable for gaps below 20 nm in size.115

5.16 The geometric dimensions of the simulated pRAM device. (in microns, not drawn to scale).......................................................................................... 117

5.17 I-V characteristics for the pRAM for different channel widths varied from 180 to 230 nm. Both gates were fixed at 0 V....................................................... 118

5.18 Transfer characteristics for the pRAM for different channel widths varied from 160 to 200 nm, with the memory gate $V_{G2}$ held at (a) zero bias and (b) negative bias. The 180 nm wide channel gives the highest on/off ratio. ... 119

5.19 (a) The memory structure showing bias conditions for simulations, (b) the bias signals applied to the memory gate and (c) a typical transient response of the drain current from which the key parameters are extracted. ........ 120
List of Figures

5.20 Drain current transient response for the pRAM with equal dimensions for both gaps (60 nm). The pulse magnitude is 10 V and the pulse width is $10^{-3}$ s. Right: Zoomed current response after the positive pulse. 

5.21 On/Off ratios and memory retention times for gap dimensions of a 1:1 ratio ranging from 20 to 100 nm. A high on/off ratio along with relatively long $t_1$ and $t_2$ are achieved for $G_1 = G_2 = 60$ nm (circled).

5.22 (a) Conduction band contour plots and (b) potential barrier heights along the dotted line (equilibrium, charged and discharged conditions). $G_1 = G_2 = 60$ nm. The conduction in the channel is controlled by the charging and discharging. The barrier height can be elevated at charged condition thus pinching off the nanogap.

5.23 (a) On/Off ratios and (b) memory retention times for combinations of gap sizes. To achieve equally long retention times for both on and off pulses, $G_1 = 20$ nm and $G_2 = 60$ nm (circled) are selected as the gap sizes.

5.24 Drain current transient response for the pRAM with $G_1 = 20$ nm and $G_2 = 60$ nm, showing a pair of substantially long memory retention times when compared to other gap dimensions. The potentials in the two gaps demonstrate that the barrier heights in the two gaps determine the leakage rate.

5.25 Drain current transient response for the pRAM with $G_1 = 20$ nm and $G_2 = 60$ nm, biased by different pulse widths. Longer pulse widths relate to longer write times, thus less efficient.

5.26 Drain current transient response for the pRAM with $G_1 = 20$ nm and $G_2 = 60$ nm, biased by different pulse magnitudes. Increasing the magnitudes achieve long memory retention times for both on and off pulses but not beyond 20 V.

5.27 (a) Electron density contour plots for $V_{G2} = 20$ and 30 V and upon restoring back to equilibrium and (b) electron density contour plots in the channel at different time moments after the positive pulse is removed; for memory biases 20 and 30 V.

5.28 (a) Drain current transient response and (b) electron density profiles along the channel at different time moments after the positive pulse is removed; for memory biases 20 and 30 V.

5.29 Conduction band contour plots showing how they change at $G_2$ for different magnitudes of positive applied bias at the memory gate.
List of Figures

5.30 Conduction band energy (E_c) probed inside the two gaps as a result of different magnitudes of applied bias at the memory gate........................................ 132

5.31 (a) Summary of the device parameters, (b) its drain current transient, and exponential fits for (c) negative pulse recovery to equilibrium and (d) positive pulse recovery to equilibrium................................................................. 133

5.32 Array structures for the pRAM. The numbers represent the number of storage regions, n. ........................................................................................................... 134

5.33 Current response for array structures. Right: t1 for n number of storage regions. .................................................................................................................. 134

5.34 Current response for oxide-filled trenches.................................................. 136
## List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1T-1C</td>
<td>One Transistor - one Capacitor</td>
</tr>
<tr>
<td>2DEG</td>
<td>Two-Dimensional Electron Gas</td>
</tr>
<tr>
<td>3D</td>
<td>Three-Dimensional</td>
</tr>
<tr>
<td>AgInSbTe</td>
<td>Silver Indium Antimony Tellurium</td>
</tr>
<tr>
<td>ALMA</td>
<td>Atacama Large Millimetre Array</td>
</tr>
<tr>
<td>BARITT</td>
<td>Barrier Injection Transit Time</td>
</tr>
<tr>
<td>BOX</td>
<td>Buried Oxide</td>
</tr>
<tr>
<td>BST</td>
<td>Barium Strontium Titanate</td>
</tr>
<tr>
<td>CAGR</td>
<td>Compounded Annual Growth Rate</td>
</tr>
<tr>
<td>CHE</td>
<td>Channel Hot-Electron</td>
</tr>
<tr>
<td>CMOS</td>
<td>Complementary Metal-Oxide–Semiconductor</td>
</tr>
<tr>
<td>CMS</td>
<td>Collision Mitigation Systems</td>
</tr>
<tr>
<td>COBE</td>
<td>Cosmic Background Explorer</td>
</tr>
<tr>
<td>CPW</td>
<td>Coplanar Waveguide</td>
</tr>
<tr>
<td>DRAM</td>
<td>Dynamic Random-Access Memory</td>
</tr>
<tr>
<td>EEPROM</td>
<td>Electrically Erasable Programmable Read-Only Memory</td>
</tr>
<tr>
<td>FET</td>
<td>Field-Effect Transistor</td>
</tr>
<tr>
<td>FLDMOB</td>
<td>Field-dependent Mobility</td>
</tr>
<tr>
<td>FMCW</td>
<td>Frequency Modulated Continuous Wave</td>
</tr>
<tr>
<td>FT-FIR</td>
<td>Fourier-Transformed Far-Infrared</td>
</tr>
<tr>
<td>GaAs</td>
<td>Gallium Arsenide</td>
</tr>
<tr>
<td>GaN</td>
<td>Gallium Nitride</td>
</tr>
<tr>
<td>GST</td>
<td>Germanium Antimony Tellurium</td>
</tr>
<tr>
<td>HEMT</td>
<td>High Electron Mobility Transistor</td>
</tr>
<tr>
<td>ICT</td>
<td>Information and Communications Technology</td>
</tr>
<tr>
<td>IMPATT</td>
<td>Impact Ionisation Transit Time</td>
</tr>
<tr>
<td>InGaAs</td>
<td>Indium Gallium Arsenide</td>
</tr>
<tr>
<td>InP</td>
<td>Indium Phosphide</td>
</tr>
<tr>
<td>IPCM</td>
<td>Interfacial Phase-Change Memory</td>
</tr>
<tr>
<td>IR</td>
<td>Infra-Red</td>
</tr>
<tr>
<td>IT</td>
<td>Information Technology</td>
</tr>
<tr>
<td>ITRS</td>
<td>International Technology Roadmap for Semiconductors</td>
</tr>
<tr>
<td>LO</td>
<td>Local Oscillator</td>
</tr>
<tr>
<td>LO</td>
<td>Longitudinal-Optical</td>
</tr>
<tr>
<td>MBE</td>
<td>Molecular Beam Epitaxy</td>
</tr>
<tr>
<td>MC</td>
<td>Monte Carlo</td>
</tr>
<tr>
<td>MITATT</td>
<td>Mixed Tunnelling-Avalanche Transit-Time</td>
</tr>
<tr>
<td>MLCS</td>
<td>Multilevel Charge Storage</td>
</tr>
</tbody>
</table>
### List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>MMIC</td>
<td>Monolithic Microwave Integrated Circuit</td>
</tr>
<tr>
<td>NEP</td>
<td>Noise Equivalent Power</td>
</tr>
<tr>
<td>NIL</td>
<td>Nanoimprinting Lithography</td>
</tr>
<tr>
<td>NRI</td>
<td>Nano-electronics Research Initiative</td>
</tr>
<tr>
<td>NVM</td>
<td>Nonvolatile Memory</td>
</tr>
<tr>
<td>ONO</td>
<td>Oxide Nitride Oxide</td>
</tr>
<tr>
<td>PC</td>
<td>Photoconductive</td>
</tr>
<tr>
<td>PCRAM</td>
<td>Phase-Change Random-Access Memory</td>
</tr>
<tr>
<td>PDA</td>
<td>Personal Digital Assistants</td>
</tr>
<tr>
<td>pRAM</td>
<td>Planar Random-Access Memory</td>
</tr>
<tr>
<td>pw</td>
<td>Pulse width</td>
</tr>
<tr>
<td>QCL</td>
<td>Quantum Cascade Laser</td>
</tr>
<tr>
<td>RAM</td>
<td>Random-Access Memory</td>
</tr>
<tr>
<td>RF</td>
<td>Radio-frequency</td>
</tr>
<tr>
<td>RTD</td>
<td>Resonant Tunnelling Diode</td>
</tr>
<tr>
<td>S.A.</td>
<td>Sense Amplifier</td>
</tr>
<tr>
<td>SAIC</td>
<td>Science Applications International Corporation</td>
</tr>
<tr>
<td>SILC</td>
<td>Stress-Induced Leakage Current</td>
</tr>
<tr>
<td>SNR</td>
<td>Signal to Noise Ratio</td>
</tr>
<tr>
<td>SOI</td>
<td>Silicon-On-Insulator</td>
</tr>
<tr>
<td>SR</td>
<td>Set-Reset</td>
</tr>
<tr>
<td>SRAM</td>
<td>Static Random-Access Memory</td>
</tr>
<tr>
<td>SRH</td>
<td>Shockley-Read-Hall</td>
</tr>
<tr>
<td>SSD</td>
<td>Self-Switching Device</td>
</tr>
<tr>
<td>TCAD</td>
<td>Technology Computer Aided Design</td>
</tr>
<tr>
<td>TD</td>
<td>Tunnel Diode</td>
</tr>
<tr>
<td>TDS</td>
<td>Time-Domain Spectroscopy</td>
</tr>
<tr>
<td>TED</td>
<td>Transferred Electron Device</td>
</tr>
<tr>
<td>THz</td>
<td>Terahertz</td>
</tr>
<tr>
<td>TDS</td>
<td>Time Domain Spectroscopy</td>
</tr>
<tr>
<td>TPI</td>
<td>Terahertz Pulsed Imaging</td>
</tr>
<tr>
<td>TUNNETT</td>
<td>Tunnel Injection Transit Time</td>
</tr>
<tr>
<td>UMS</td>
<td>United Monolithic Semiconductors</td>
</tr>
<tr>
<td>UTC-PDs</td>
<td>Uni-Travelling-Carrier Photodiodes</td>
</tr>
<tr>
<td>ZnO</td>
<td>Zinc Oxide</td>
</tr>
</tbody>
</table>
1

Introduction

Today, information technology (IT) is not only the largest component of the world’s economy but it changes and shapes all walks of human life. The unprecedented growth of the IT industry has largely been due to the exponential increase in the performance of the semiconductor chips that are at the heart of all modern electronics [1]. The key component on these chips is the complementary metal-oxide–semiconductor (CMOS) field-effect transistor (FET), and the ability to scale these devices to ever-smaller dimensions has been the primary driver of this increased performance. For over 30 years, the industry has been able to pack twice as many FETs onto a chip every 18–24 months, in what has come to be known as "Moore’s law" [2]. This has resulted in an exponential increase in the information processing capability per unit area on the chip, or more importantly, per US dollar. This has meant not only that existing chip-based products get faster and/or cheaper each year, but also expanded the number of products that use semiconductor chips to increase functionality, from toasters to mobile-phones to supercomputers.

The rules for FET scaling that have enabled this revolution were outlined by Dennard et al. in the early 1970s [3]. The key insight was that if all of the critical dimensions of the FET, along with the operating voltage, were reduced by the same factor, the speed of the FET would go up while the area and power would go down, so that the power density remained constant. However, in recent chip generations, the ability to scale the voltage has become limited due to the difficulty to maintain performance giving rise to circuits whose power density is
1.1 The Terahertz gap

Figure 1.1 Intel CPU trends. Moore’s Law is expected to continue for at least a few more years, but raw clock cycles have met their limit. Any future performance gains are going to be accomplished in fundamentally different ways for at least the next couple of processor generations. Most current applications will no longer benefit without significant redesign [4].

...approaching 100 W/cm², Fig. 1.1. The costs incurred from managing this heat load are significant and have forced designers to effectively spread the heat load more equitably over the entire chip thereby giving rise, for example, to multi-core processors. By distributing the heat load over the chip, designers have temporarily fore-stalled the heat management problem but they have not circumvented it permanently.

The physical scaling of FETs is bound to be problem prone at channel lengths in the order of 5 nm due to increasingly unpredictable behaviour arising from electron tunnelling and over-barrier transitions [5]. While this still leaves a lot of room for scaling over the next 10 – 15 years, the twin challenges of heat management and quantum effects are forcing many semiconductor scientists and engineers to re-evaluate the role of switching electron charge in FETs as the salient information processing technology to extend much beyond 2020, when it is thought
1.1 The Terahertz gap

that the physical limits of scaling may be reached. Moreover, the power concern is not unique to silicon (the semiconductor used in CMOS FETs) but would in fact affect FETs in any material, including carbon nanotubes, organic molecules and others. While changing materials might improve FET operation for a generation or two (which certainly might be worthwhile), new devices need to be found to allow long-term continued scaling; where scaling should be understood in the most generic sense of increasing computational performance per unit area/dollar in each subsequent generation.

The International Technology Roadmap for Semiconductors (ITRS) Emerging Research Device Technical Working Group began to study the challenge presented by power density for future scaling in the early 2000s while the Nanoelectronics Research Initiative (NRI) was handed the task of demonstrating novel computing devices capable of replacing the CMOS FET as a logic switch in the 2020 timeframe. The goal for these devices is to show significant advantage over ultimate FETs in power, performance, density, and/or cost to enable the semiconductor industry to extend the historical cost and performance trends for IT. To meet these goals, the NRI has focussed research on devices utilising new computational state variables and switching mechanisms. In addition, the NRI is interested in new interconnect technologies and novel circuits and architectures, including nonequilibrium systems, for exploiting these devices, as well as improved nanoscale thermal management and novel materials and fabrication methods for these structures and circuits. Finally, it is desirable that these technologies be capable of integrating with CMOS, to allow exploitation of their potentially complementary functionality in heterogeneous systems and to enable a smooth transition to a new scaling path.

It is interesting to note that the current FET was developed when vacuum tubes and mechanical switches were having similar reliability and power challenges of their own. Therefore, it is reasonable to anticipate that a path will soon exist for mapping an entire existing high-performance microprocessor design into proposed new state variables and technologies. These approximations will provide the first credible glimpse into what conventional computing will look like after CMOS; alternative architectures supported by new switches, however, may enable even more profound advances. It is this sense of purpose and motivation that sets the direction
1.1 The Terahertz gap

for the work produced in this thesis. Planar nanoelectronic devices having ultrafast
switching speeds are investigated targeting two niche sectors in the electronics
industry: terahertz and memory. These sectors are identified to actively demand
radical changes to their existing device nomenclatures to make any further progress.
An overview considering these needs follows.

1.1 The Terahertz gap

Over the last century, physicists and engineers have progressively explored and
conquered the electromagnetic spectrum. Starting with visible light, techniques have
been developed in generating and detecting radiation at both higher and lower
frequencies. And at each successive region of the spectrum that has been colonised,
the technology required to exploit the radiation has always been developed. X-rays,
for example are routinely used to image hidden objects. Near-infrared radiation is
used in fibre-optic communications and in compact-disc players, while microwaves
are used to transmit signals from mobile phones.

However one part of the electromagnetic spectrum that has steadfastly resisted
advancement is the terahertz (THz) region, which ranges from frequencies of about
300 GHz to 10 THz (10×10^{12} Hz). This corresponds to wavelengths of between 1
and 0.03 mm and lies between the microwave and infrared regions of the spectrum
(Fig. 1.2). However, the difficulties involved in making suitably compact THz
sources and detectors has meant that this region of the spectrum has only begun to be
explored thoroughly over the last decade.

![Figure 1.2 Representation of the electromagnetic spectrum illustrating the THz gap relative to the microwave and infrared.](image-url)
1.1 The Terahertz gap

Much of the recent interest in THz radiation stems from its ability to penetrate deep into many organic materials without the damage associated with ionising radiation such as X-rays. Additionally, THz radiation is readily absorbed by water, thus it can be used to distinguish between membranes with varying water content. These properties can lead to very appealing applications in biomedical imaging [6].

THz radiation can also penetrate poor weather, dust and smoke far better than infrared or visible systems. For this reason defense and security applications are also envisaged. Furthermore the ability to penetrate dielectrics such as windows, paper, clothing and in certain instances even walls, has opened up the opportunity to reveal concealed explosives, metallic and non-metallic weapons (such as ceramic, plastic or composite guns and knives), chemical weapons and other threats hidden in packages. THz detection is also a very interesting field for astronomers: much of the photons generated in the Big Bang fall into the sub-millimeter and far-infrared region of the spectrum. Nondestructive testing of integrated circuits is also a future area of application [7]. BCC Research (one of the leading semiconductor industry information sources) estimates the market for THz radiation devices totalled $83.7 million in 2011. This market will grow to $127 million in 2016. The diversification of the THz market is expected to accelerate after 2016, and the total market should reach $570 million by 2021, a compounded annual growth rate (CAGR) of 35% from 2016 to 2021 (Figs. 1.3a and b) [8].

A particularly intriguing feature of THz radiation is that the semiconductor devices that generate radiation at frequencies above and below this range operate in completely different ways. At lower frequencies, microwaves and millimeter-waves can be generated by electronic devices such as those found in mobile phones. At higher frequencies, near-infrared and visible light are generated by optical devices such as semiconductor laser diodes, in which electrons emit light when they jump across the semiconductor band gap. Unfortunately neither electronic nor optical devices can conveniently be made to work in the THz region. Developing a THz system is therefore a challenging business because it involves working in a region where established solid-state technologies fail.
1.1 The Terahertz gap

Nevertheless, motivated by the potential applications, researchers strive to bridge the gap and develop breakthrough systems operating in the THz frequency range of the electromagnetic spectrum [9, 10]. Although there has been significant progress in being able to detect THz radiation, realising effective sources in the same regime is posing the toughest challenge, particularly since current systems are in one way or another limited by their size, cost, output power, efficiency or required temperature of operation [11]. For example, the quantum cascade laser shows exceptional promise, but operation in the important sub-1-THz regime has not been achieved, and cryogenic cooling is required [12]. Because of its importance to the motivation of this thesis, the wide array of THz sources is more elaborately reviewed in Chapter 2 to gain background knowledge of their respective merits and demerits.

Figure 1.3 Global market for THz radiation devices and systems through 2021, (a) by type of system and (b) by application. Source: BCC Research.
1.2 Semiconductor memories

Memory can be defined as the ability to store the state of a system at a given time, and access such information, or part of it, at some later time [13]. This state could be the resistance, capacitance, inductance, spin polarisation, the doping profile, or some other physical characteristic of the system [14]. These characteristics can function on their own or in combinations leading to memory in different materials and systems.

Over the last several decades, processors have improved much faster than storage components [15]. Memory has therefore become the ultimate limiting factor of a system’s performance; and the convergence of consumer, computer, and communication electronics has exponentially increased the need for increased memory size [16]. In the future, a conglomeration of integrated intelligence is destined to lead to more issues. Pressure is certain to be mounted on cost, performance optimisation, flexibility and integration density. From this perspective, novel solutions provided by emerging memory technologies are expected to meet the cost, bandwidth, and power efficiency requirements of future memory systems.

The technologies that have been developed over the years in realising memory or data storage can be divided into semiconductor types (MOS, bipolar, and charge-coupled devices) and moving media types (magnetic disk, optical disk) which require mechanical equipment operation [17]. Compared to disk devices, semiconductor memories exhibit superior characteristics in terms of cost and performance [18]. Furthermore, due to the advantages of MOS technology in device manufacturability and miniaturisation, most ultralarge-scale integration (ULSI $\geq 10^7$ transistors on a chip) memory circuits are made at the present time using MOS memories. The most important device for MOS memory technology today is the metal–oxide semiconductor (MOS) field-effect transistor (FET) [19, 20].

Ideal memory devices ought to be low-cost, nonvolatile, able to retain data without external power, have random access that features read–write access to any individual data location, high speed, high density, low power consumption, be easy to test, highly reliable, and compatible with established industrial manufacturing routes [17]. Unfortunately a single type of memory device having all of these characteristics remains unfeasible to this day. Instead, the MOS memory devices that are used as the best alternatives in today’s electronics systems can be classified in
1.2 Semiconductor memories

terms of their volatility and access speed as volatile random-access memories (RAMs) and nonvolatile non-RAMs. The leading technologies in these two sectors are the Dynamic Random-Access Memory (DRAM) and the flash memory, respectively.

Although DRAM and flash are unique to their respective applications, competition has been rife in the race to gather a bigger share of the semiconductor market. It is no surprise that the flash memory market has quickly gained dominance over the DRAM market that has held its own for over four decades, without which, the computer industry would be non-existent. The rise of the flash memory sector, Fig. 1.4, has been fuelled largely by the escalating demand in consumer electronics. This is because flash memories possess two qualities that are desired by present day electronic devices: mobility and miniaturisation. For example, mobile phones, a major application for flash memory, require data storage to save and store frequently called numbers and perform other convenient functions for which a traditional hard drive would prove impractical; such information would be erased each time the phone was switched off. Because flash memory is small and reliable; and its memory is nonvolatile, numerous applications not practicable with traditional data storage technology are emerging. Popular consumer electronics such as USB flash memory drives, DVD players, digital cameras, MP3 players, personal digital assistants (PDAs), global positioning systems (GPS), etc., would not have been possible without flash memory [21].

Figure 1.4 Sales figures showing the growth in the flash memory unit shipments overtaking that of DRAM. Source: Market research firm, IC Insights Inc.
However the separate superiority held by these two giant technologies, that is the flash and DRAM, is gaining a balance as applications such as smartphones and tablet personal computers demand both to be integrated within the same device. The challenge is not so much of just the volatility anymore, but one that demands low power or in some cases no power at all, portability, high speeds, and most importantly, high integration densities. Cost is always another economically non-ignorable factor. To this end, device mechanics have met their scaling limit owing to reasons unique to their mechanisms. For DRAM, scaling above gigabit densities (over 8 Gbits) is extremely difficult due to the inherent constraints of low-leakage access transistors and the need for large storage capacitance [17]. For flash, the tight spacing, floating gate interference and the need for sufficient gate control (gate coupling ratio) have also ruled out the continuation of the conventional floating gate device below approximately 32 nm node [22]. To overcome these technological constraints, new memory concepts are needed.

The developments in this area are also driven by emerging large-area microelectronic applications, such as rollable displays [23], electronic paper [24], contactless identification transponders [25, 26], and smart labels [27]. Thus, new architectures would be necessary to deliver DRAM and flash memories for the future. They must be drop-in compatible with existing memory and have favourable cost, power, reliability, and performance characteristics.

### 1.3 Computational electronics

As semiconductor feature sizes shrink into the nanometer scale regime, even conventional device behaviour becomes increasingly complicated as new physical phenomena at short dimensions occur, and limitations in material properties are reached [28]. In addition to the problems related to the understanding of actual operation of ultra-small devices, the reduced feature sizes require more complicated and time-consuming manufacturing processes. This fact signifies that a pure trial-and-error approach to device optimisation will become impossible since it is both too time consuming and too expensive. Since computers are considerably cheaper resources, simulation is becoming an indispensable tool for the device engineer. Besides offering the possibility to test hypothetical devices which have not (or could
not yet been manufactured, simulation offers unique insight into device behaviour by allowing the observation of phenomena that cannot be measured on real devices. The modelling and simulation of semiconductor devices have played a central role in the semiconductor miniaturisation process. It is related to, but usually separate from process simulation, which deals with various physical processes such as material growth, oxidation, impurity diffusion, etching, and metal deposition inherent in device fabrication leading to integrated circuits [29].

With the development of new fabrication techniques and device concepts during the last five decades to obey Moore’s Law [30], computational electronics [31, 32] has been improving in parallel the simulation tools and models to predict and optimise the performance of next generation devices. These efforts have allowed an important reduction in design time and cost. As a consequence, there is a wide spectrum of available approaches to describe the behaviour of a considered technology. Different solutions from classical to full quantum models can be considered depending on the needed accuracy, the computational resources and the available time to perform the simulations [33-37]. Numerical simulations of semiconductor devices pursue different goals. Among many others, the following can be cited [38]:

(i) The behaviour of a particular realistic device can be predicted, for example, calculating the current obtained in their terminals when a given bias is applied to them. This is useful for improving the device performance by changing the technological structure of the device or to understand the dependencies and limiting physical mechanisms in the device/circuit performance (e.g. effects of noise, limits on frequency/gain, trap effects, effects of geometry, etc.)

(ii) Transport properties of charge carriers can be studied, e.g. mobility, diffusion coefficient, velocity overshoot, in different materials at different conditions of temperature, quantisation, strain, crystallographic orientations, and many others.

(iii) The physics underlying device behaviour can be understood and explained much better than it can be experimentally because known models are used. In this context, simulation can be used to develop and
improve device compact models which are essential in the design process of new integrated circuits.

While particle-based device simulation methods, such as the Ensemble Monte Carlo technique [39, 40] have been used for well over 30 years, the relatively new physically-based device simulation is growing in popularity through commercially available products like ATLAS by SILVACO, inc. Physically-based device simulators predict the electrical characteristics that are associated with specified physical structures and bias conditions. This is achieved by approximating the operation of a device onto a two or three dimensional grid, consisting of a number of grid points called nodes [41]. The transport of carriers through a structure is simulated by applying a set of differential equations, derived from Maxwell’s laws, onto this grid. Thus the electrical performance of a device can be modelled in DC, AC or transient modes of operation. Physically-based simulation provides three major advantages: it is predictive, provides insight, and captures theoretical knowledge in a way that makes this knowledge available to non-experts. Physically-based simulation is different from empirical modelling. The goal of empirical modelling is to obtain analytic formulae that approximate existing data with good accuracy and minimum complexity. Empirical models provide efficient approximation and interpolation. They do not provide insight, or predictive capabilities, or encapsulation of theoretical knowledge. Physically-based simulation is an alternative to experiments as a source of data.

Physically-based simulation has become very important for two reasons. First, it is almost always much quicker and cheaper than performing experiments. Second, it provides information that is difficult or impossible to measure. The drawbacks of simulation are that all the relevant physics must be incorporated into a simulator, and numerical procedures must be implemented to solve the associated equations. This is the reason for the high accreditation given to the SILVACO package in the semiconductor fraternity; because of its robust and well-developed physical models. Users of physically-based device simulation tools generally need to specify the problem to be simulated [41], for example, in ATLAS, users specify device simulation problems by defining:
1.4 Motivation and aims

- The physical structure to be simulated
- The physical models to be used
- The bias conditions for which electrical characteristics are to be simulated.

1.4 Motivation and aims

The continued existence of the THz gap persists largely due to the scarcity of compact, cost-effective, tunable and efficient THz sources that are able to operate at room temperature. Numerous novel devices and materials have been the subject of intense research to meet this demand as it would enable a multitude of applications and make huge progress in a number of sectors, particularly the medical field where imaging and sensing of tissues and cells are extremely critical and needed. The most promising solution thus far has been to utilise solid-state devices, which offer the best possibilities of integration with other electronic or optoelectronic devices within a single chip [42]. Planar nanostructure devices based on two-dimensional electron gases (2DEGs) in semiconductor heterostructures, e.g., high electron mobility transistors (HEMTs), have demonstrated emission at THz frequencies [43-45]. More recently, a novel planar unipolar device, namely, self-switching device (SSD) [46, 47] has demonstrated room temperature electrical rectification at 1.5 THz [48] with excellent noise-equivalent power [49]. The key feature of the SSD is the ability to construct an array of devices connected in parallel using only a single lithography step. Due to the truly planar nature of the device, it eliminates the need for interconnection layers which may introduce parasitic elements. This not only enables rectification at ultra-high speed, but also makes the whole fabrication process simpler, faster and inexpensive. Theoretical simulations have predicted THz emission at 0.13 THz in InGaAs [50] and up to 0.6 THz in GaN-based devices [51] by means of Gunn oscillations. So far, however, no systematic studies on the device geometrical parameters have been reported, which is one of the objectives of this work. The motivation arises from being able to enhance Gunn oscillation output power by a factor at least equal to the number of SSDs or even more if there is any phase locking of the oscillations in individual nanochannels. Therefore the prospect of designing individual SSDs with different nanochannel widths/lengths to generate
1.4 Motivation and aims

different oscillation frequencies on a single chip is groundbreaking. The feasibility of switching on and off each individual SSD separately will also make it convenient for THz spectroscopy measurements. A detailed documentation of the advantages is reported in Chapter 2. In addition, previous studies showed that the presence of surface charges on semiconductor-insulator interfaces has a strong influence on the DC characteristics of the device, particularly the threshold voltage. Therefore, the influence of these surface charges on the device’s emission properties is also closely examined.

Over and above the emission characteristic, the SSD has been demonstrated in a number of roles such as a diode, RF rectifier and THz detector with both high and low mobility materials proving its versatility in a wide range of applications. In this regard, logic circuits based on SSDs have been demonstrated [52]. The memory effect found in the SSD by observing hysteresis in its $I-V$ is a truly unique discovery that adds another new dimension to the functions of such a simple device. This property, however, has been given much less attention in comparison to the others thus far. Led by the intensified demand for scalable, low-cost memories to deal with the high volume of consumer electronics and future technologies, the development of the SSD memory feature was deemed as a necessary and timely step. The planar architecture of the device will certainly meet the urgent requirement for high density integrated memories. Moreover, the ability to use low-cost materials to construct single plane devices and circuits is exciting for emerging applications such as organic, printed and flexible electronics. For this, a novel planar device is conceptualised and studied in detail for memory applications.
2

Background

2.1 Terahertz (THz) technology

The terahertz (THz) region, typically referred to as the frequencies from 0.1 to 10 THz (Fig. 2.1), has been the least explored region of the electromagnetic spectrum for a long time. However, the accumulating number of potential applications for THz waves in recent years; have seen the emergence of innovative and improved technologies for generation and detection of these waves.

The principle of applications for THz technology can be classified under two categories: sensing and communications. THz time-domain and frequency-domain spectroscopy methods provide sensing capabilities for many research fields including biology, pharmacy, medical science, material science, astronomy,
2.1 Terahertz (THz) technology

Environment monitoring, industrial non-destructive evaluation and security. Information and communications technology such as wireless communication, high-speed data processing and satellite communication also benefit from THz technology. The synergy of these research areas promise a futuristic outlook, enabling applications such as biometrics where recognition based on unique human physical or behavioural traits are possible with THz cameras, sensor networks and selective communication.

In the following sections, these applications will be described in greater detail followed by THz generation techniques citing their merits and demerits. A separate section is devoted to the use of solid-state devices, particularly the Gunn diode, since they provide the most practical solution to meeting the requirement of reducing size and cost. Finally, a historical backdrop to the self-switching device (SSD) is provided since it inspires much of the work produced in this thesis.

2.1.1 THz applications

Radio astronomy

Radio astronomy has been a major driver for the development of THz technology [54], [55], for many years. The fundamentally important objective has been escaping the attenuation of the Earth’s atmosphere. This has led to the development of a significant range of missions, entailing the observation of both interstellar and extragalactic regions of space. Past missions such as the NASA Cosmic Background Explorer (COBE) and the development of Planck and Herschel have provided vast information on the universe in the sub-millimetre/THz spectral region. Utilising THz waves stems from their ability to see through dust clouds and surrounding for example, star forming regions [56]. Observations over a wide range of frequencies in this spectrum test different cosmological theories of the evolution of the Universe (Fig. 2.2). Also, the spatial extent of star formation phenomena is quite varied (from arc-minutes to degrees) and, consequently, it is necessary to image wide-field regions of space with high spatial resolution. The provision of instrumentation that is simultaneously capable of performing wide-field and high spatial resolution measurements is therefore of great interest to astrophysicists.
2.1 Terahertz (THz) technology

In keeping up with the rapidly evolving landscape for space applications, device technology is almost always required to meet the challenges without delay. In recent times, a new generation 'Microwave Limb Sounder' installed on NASA's 'Earth Observing System has been developed by NASA to monitor additional molecular species (such as hydroxide ion OH) at frequencies as high as 2.5 THz. The Atacama Large Millimetre Array (ALMA) is an international astronomy facility in Northern Chile, with the ability to detect waves close to 1 THz frequencies passing through the earth’s atmosphere [58]. The monitoring could help probe the invisible dark universe, including newborn galaxies as far as 13 billion light years away, the birth of a new solar system, or extraterrestrial organic molecules. Although, ultra-low noise cryogenic superconductive junctions would be used for the mixers in ALMA, local oscillator (LO) powers would still be provided by GaAs multiplier diodes [56]. Such diode system was used to detect space shuttle insulation foam defects. The system adequately identified defects by a using 12 mW, 0.2 THz Gunn diode oscillator [59].

Security

Never has the threat to security been so critical than in recent years owing to the changes in the geopolitical climate; requiring a multifold need for reliable mm-wave imaging systems at sensitive locations globally. In particular, the airport and sea port
2.1 Terahertz (THz) technology

security markets are expected to grow exponentially [60]. Every explosive and narcotic has a distinct signature in its THz spectra, making THz spectroscopy valuable for security applications. These signatures allow the identification of many chemicals through their transmission spectra. As waves below 3 THz can generally pass through envelopes, materials can be identified using THz multispectral images and component spatial-pattern analysis without having to open the mail, Fig. 2.3a [61].

Figure 2.3 (a) Terahertz narcotic detection: THz image (top) and photograph (bottom) of specimens under inspection. The three kinds of powder hidden in the envelopes are distinguished using THz-TDS by their own THz signatures. Image courtesy of RIKEN, Japan. (b) ThruVision’s passive imaging security system detects terahertz radiation emitted from objects hidden on the body. Source: ThruVision Ltd.

The development of a focal-plane array for THz camera has also been of great technological importance. The camera images objects through fog and smoke, regardless of the background illumination. For example, a passive sub-THz camera made up of a Schottky barrier diode array with a silicon photonic-bandgap crystal and a heterodyne detection method operating at a frequency of about 0.5 THz is being used in airport security [58] (Fig. 2.3b). An active THz camera would suit even more applications, including biometrics.

Some security applications may require remote detection over a significant distance, but THz active sensing becomes unsuitable due to its sensitivity to the environment. Recently, remote-controlled THz generation using air plasma has
2.1 Terahertz (THz) technology

raised the level of interest. An intense femtosecond laser beam is focussed near the object to generate an ambient plasma, which produces THz waves, and the reflection from the target is detected in a similar way to THz-Time Domain Spectroscopy (THz-TDS) [62]. The method of THz-TDS is contained in 2.1.2.

Biomedical and Pharmaceutical

Knowledge on the biological makeup of the human body is absolutely essential to progress in medicine, be it for cures of diseases or developing drugs. Thus the non-invasive nature of THz waves aids the study of biomolecules, medicines, cancer tissue, DNA, proteins and bacteria tremendously. There have been several studies on the diagnosis of cancer using THz spectroscopy and imaging [63]. The absorption of THz waves is sensitive to polar molecules, such as water. As such, cancer tissue which has different hydration levels from normal tissue, can be detected by means of unusual reflections. As the THz transmittance of ice is much higher than that of water, cancer cells can also be detected in frozen tissue. Similarly, caries in human teeth can be detected at an early stage in the enamel layers, Fig. 2.4. If decay can be detected early enough, it is possible to reverse the process without the need for drilling by the use of either fissure sealing or remineralisation. Instruments such as the Terahertz Pulsed Imaging (TPI™) system built by Teraview Limited are already in commercial use for these purposes.

Figure 2.4 TeraView’s Terahertz Pulsed Imaging (TPI™) is able to distinguish between the different types of tissue in a human tooth; detect caries at an early stage in the enamel layers of human teeth and monitor early erosion of the enamel at the surface of the tooth. (a) X-ray image. (b) THz image. Source: TeraView Ltd.
2.1 Terahertz (THz) technology

Figure 2.5 THz absorption spectra of barbital. (a) THz-TDS. (b) FT-FIR. The polymorph sample (form B) was obtained by heating original barbital (form A) at 160 °C for 30 minutes [65].

Another application is the classifying of polymorphs in medicine [64]. Figure 2.5 shows THz-absorption spectra of barbital (a hypnotic drug), which changes its form on heat treatment [65]. The differences are far more distinguishable in the THz time-domain spectroscopy spectra compared to the Fourier-transformed far-infrared (FT-FIR) data.

Information and communications

The information and communications technology (ICT) sector is pivotal in today’s modern society where the market for it is enormous. Information flow and carrier frequency of signals are continually growing. Fibre-linked optical communication is attaining data rates above a terabyte per second while wireless communication data rates remain relatively low. THz-ICT would greatly benefit this void by enabling high-performance wireless connections for applications such as: communication in rural areas; communication between buildings during disasters; high-vision data delivery for telemedicine and outdoor entertainment; hot spots for movies; and even data transmission in media and sports.

The advent of sub-THz-wave generating uni-travelling-carrier photodiodes (UTC-PDs) is helping realise such possibilities, such as point-to-point line-of-sight 'last mile' communication links by Bridgewave Communications Inc. and E-Band Communications Corp. [66]. These 120-GHz-band millimeter-wave wireless links
2.1 Terahertz (THz) technology

operate at over 10 Gbit per second [67] which is fast enough to download a typical movie in seconds. The offered per link cost is approximately one tenth of the commercially available MMIC chip-sets fibre links. Higher frequencies could well achieve higher bandwidths and narrow beam widths integral to indoor wireless communications.

The delivery of THz waves is important for both ICT and sensing applications. Approaches to this include plastic fibre and polycarbonate waveguides. Interestingly, bare metal can transport THz pulses along its surface with virtually no dispersion and low attenuation [68]; this is explained in Sommerfield’s description of an electromagnetic wave propagating along the surface of a cylindrical conductor. This flexibility is useful for making beam splitters. Recently, special photoconductive antennas with radial symmetry have been developed to enhance the coupling coefficient of the THz waves to cylindrical-wire waveguides [69].

Radar

Radar is one of the most commonly used millimetre-wave systems, whose spatial resolution rises with frequency due to narrowing of the achievable beamwidth. The automotive industry has been mainly utilising the mm-wave radar systems. A recent study by TRW Automotive® 2011 showed that road accidents involving heavy vehicles were reduced by 70 percent due to the installation of collision warning and avoidance systems on vehicles [70]. Radar systems using both Pulse Doppler and frequency-modulated continuous-wave (FMCW) modes, operating at 77 GHz have been effectively installed and used in top of the range automobiles for over a decade. Future deployment of Collision Mitigation Systems (CMS), which would initiate various safety actions such as seatbelt pre-tensioning and airbags deployment upon detection of a probable collision has already begun [70]. Another significant FMCW radar application includes its utilisation in a factory for vibration and displacement measurements. In this regard, coherent radar operating at 37.5 GHz with 50 mW power has been developed and tested, which would be an alternative to traditional piezoelectric sensor measurements [71].
2.1 Terahertz (THz) technology

2.1.2 Time-domain spectroscopy

The first and by far the largest wave of THz activity is centred on the THz time-domain spectroscopy (THz-TDS) approach, which was pioneered in the late 1980s by groups at AT&T Bell Laboratories [72] and IBM’s T. J. Watson Research Center [73]. THz-TDS, or what later became to be known as “T-rays” when it was applied to imaging [74, 75], utilises relatively simple and inexpensive electronic technology and ultra-fast laser pulses (i.e., to both generate and detect the radiation) to realise very broadband THz sensing and imaging [53]. As shown in Fig. 2.6, the main laser beam is split into two beams by a beam splitter. One beam (called the pump) is focussed onto a PC antenna (emitter) to produce THz radiation, which then passes through the sample to be analysed.

![Figure 2.6 Time-domain spectroscopy (TDS) setup [77].](image)

The transmitted beam, which carries information about the sample, is then focussed onto a detector, either a PC antenna or an EO crystal. The other beam (called the probe) is also focussed onto the detector for the detection process. The output signal from the detector is time-dependent on the magnitude and the sign of the THz. Therefore, by changing the arrival of the probe beam (time delay) using a computer-
2.2 THz generation

Conventionally, THz radiation is generated by heat sources as a segment of the black body radiation. The energy of a 1 THz photon is merely 4.1 meV, which corresponds to a temperature of 48 Kelvin. Owing to its natural low intensity and the complexities in distinguishing the THz signals from the thermal background, various techniques have been established to generate coherent THz radiation of a higher intensity. The ideas are generally developed from the same techniques as infra-red (IR) or microwave radiation above and below the THz frequency range, that is, a frequency-down conversion from the optical, higher frequency side and a frequency-up conversion from the electronic, lower frequency side as illustrated in Fig. 2.7.

Figure 2.7 Generation of THz radiation using optical and electronic techniques [77].
2.2 THz generation

2.2.1 Optical approaches

Optical approaches to generation focus on applying existing technology with nonlinear materials. Examples include using either a femtosecond laser pulse [78, 79] or the heterodyning (photomixing) of optical/infrared lasers [80] to illuminate a photoconductive-antenna combination.

Heterodyning (Photomixing)

This method, also known as optical heterodyne downconversion, combines two CW laser beams that have slight difference in their frequency to generate CW THz radiation from a photoconductive (PC) switch. Combining the two beams, which have energy above the band gap of the photoconductive material, generates beatings with a frequency equal to the difference between their frequencies; this can be tuned so that it lies in the THz range. The generated beatings can modulate the conductance of the photoconductive material; hence, by applying an electric field through metal contacts deposited on the PC material, a current modulated at the THz frequency can be generated and fed to a planar antenna to generate THz radiation. Achieving this provides a tunable THz source with a broad tuning range using tunable laser sources. An example of this is shown in Fig. 2.8 for the case of heterodyned lasers, which produce continuous wave radiation for the operation part of the duty cycle at a narrow band frequency. Also shown is the hemispherical lens, which is often used to improve radiation, coupling from the relatively high dielectric constant substrate to free space.

Photoconductive switching

Alternatively, illuminating the photoconductor using a single femtosecond pulse creates a THz wave packet of wide bandwidth but short duration. The choice of illumination therefore depends on the application. However, both methods are competent of creating relatively high power mW operation at room temperature. Cost and bulkiness are the major drawbacks of these methods. Both the lasers and the optical benches needed to align them are costly although efforts are underway to
2.2 THz generation

![Figure 2.8 Schematic of a heterodyne THz source. Top: Beams from two lasers are combined and the resultant THz beat focussed onto the antenna which radiates the THz waves through a hemispherical lens to better couple it to free space. Bottom: Spiral antenna on a photoconductive (PC) GaAs substrate. A bias is placed between the photomixer electrodes so that electrons and holes generated by the THz beat are attracted to opposite sides [82].](image)

Integrate all the components of the heterodyne system into a single semiconductor stack by use of multi-mode semiconductor lasers [81].

Molecular gas

The principle of these THz gas lasers is based on an injection of the high power CO$_2$ laser into a cavity containing molecular gases, such as CH$_3$F, CH$_3$OH, NH$_3$ and CH$_2$F$_2$ as a gain medium. The THz radiation is then generated from the rotational transitions of the molecules. The lasing frequency of these lasers depends on the filling gas [83] and the power level is in the range of 1-100mW. For example, 50 mW THz radiation at a frequency of 2.5 THz was generated from a molecular gas laser that had methanol as an active medium pumped by a 20W CO$_2$ laser [84]. One of the downsides to this method is the discrete lines formed in the frequency spectrum by the rotational transitions of different molecular gases.
2.2 THz generation

Quantum Cascade Laser (QCL)

Although there are no suitable bulk semiconductors with an appropriate band gap, semiconductor heterostructures can be layered to create a series of quantum wells (or superlattice) engineered to have the required transitional energy for THz emission. These systems, called Quantum Cascade Lasers (QCLs), have made swift progress in recent years [85-87]. The THz waves are emitted by means of electron relaxation between subbands of quantum wells; for example, between several few-mm-thick GaAs layers separated by Al$_x$Ga$_{1-x}$As barriers, whose emission blocks are serially connected (Fig. 2.9a).

An example of the potential and band structure of a QCL is shown in Fig. 2.9b. Injection layers and active layers are alternately repeated many times with each layer comprising of several potential barriers, such that a miniband is created. THz waves are generated in the active region by electron relaxation from band 4 (green) to band 3 (blue), followed by the relaxation from band 3 to band 1 (red) by means of longitudinal-optical (LO) phonon mode scattering, $h$ is Plank’s constant and $v$ is the frequency.

![Schematic of a semiconductor source of THz waves based on a QCL design.](image)

![Conduction band profile of a GaAs-AlGaAs quantum-well-based QCL.](image)

Initially limited to cryogenic temperatures due to thermal noise affecting the intersubband transitions, continuous wave operation has been demonstrated at room temperature at infrared wavelengths [88]. Simulations have shown that room
2.2 THz generation

temperature operation could also be possible at THz wavelengths [89]. Frequencies attained previously tended to be in the high THz region, although recent operations have managed down to 1.39THz [12] with predictions of operation down to 1 THz [90].

QCLs have the advantage of relatively high power (tens of milliwatts) and continuous wave operation for a reasonable duty cycle. Created in a single heterostructure, they also have the potential for inexpensive fabrication. The disadvantage though, is that they still require cryogenic cooling and operate at a single, narrowband frequency that is fixed for any given heterostructure wafer.

2.2.2 Electronic approaches

High Electron Mobility Transistors (HEMT)

Electronically, THz generation is most commonly accomplished by the high electron mobility transistor (HEMT). The cut off frequency, $f_c$ can be approximated by

$$ f_c \approx \frac{1}{2\pi \tau} \quad (2.1) $$

where $\tau$ is the electron transit time across the transistor. Therefore, smaller transistors and higher mobilities increase the cut off frequency. However, both are limited. At room temperature, phonon scattering restricts mobility, while there are practical limits to transistor sizes for commercial fabrication. The best conventional HEMTs therefore top out at a few hundred gigahertz [91], although an emerging extension to HEMT operation is plasma wave generation of radiation. With the high electron densities that can be produced in source drain channels of operational HEMTs, the electron-electron interactions can become substantial. As such, they behave as a frictionless fluid with the same governing equations as hydrodynamics [92]. This sets up the instability conditions where plasma waves can oscillate in the channel. With high propagation velocities, these have been shown to produce radiation up to 4.5 THz [44, 93].
2.2 THz generation

Free Electron Lasers (FEL) and Synchrotrons

At the opposite end of the size scale, very high power THz emission with tunable frequencies is available from both synchrotrons and free electron lasers. The free electron laser passes a beam of electrons in ultra-high vacuum through an alternating set of magnets as shown in Fig. 2.10. Under the influence of the magnets, the moving electrons oscillate, generating radiation. The electromagnetic field stimulates the emission causing lasing to occur. The frequency of operation is dictated by the velocity of electrons between the magnets.

The synchrotron also uses a beam of electrons in high vacuum, but generates radiation at kinks in the electrons path as they are deflected by magnets around a circular path. The acceleration by the magnets at these points creates an incoherent but powerful source, again with frequency dictated by the electron velocity. Both synchrotron and free electron laser sources can produce kilowatts of power at variable frequencies. However, the construction and operational cost mean that only a small number exist worldwide and facilities are only available at a national level with beam time difficult to obtain [94].

![Figure 2.10 Schematic of a free electron laser](image)

2.2.3 Solid-state devices

Microwave frequencies can be generated using various two-terminal solid-state active devices. The most commonly used devices include transferred electron or Gunn diodes (TEDs) [96], Esaki tunnel diodes (TDs) [97], resonant tunnelling
2.2 THz generation

diodes (RTDs) [98] and transit-time diodes. The transit-time diodes include Impact Ionisation Transit Time (IMPATT) [99], Barrier Injection Transit Time (BARITT) [100], Tunnel Injection Transit Time (TUNNETT) [101] and Mixed Tunnelling-Avalanche Transit-Time (MITATT) [102, 103] diodes. All of these devices display the property of negative differential resistance (NDR), Fig. 2.11.

Solid-state sources, like oscillators and amplifiers, are typically limited in frequency due to the transit time of carriers through semiconductor junctions, which causes high frequency roll-off. But, they are rugged, compact and can operate CW at room temperature with a relatively narrow linewidth ($10^{-6}$). A CW power of about 100 mW can be gained around 100 GHz [104]. Owing to their compactness, the scope of application of these sources is rapidly growing.

![Figure 2.11 I-V of a device showing negative differential resistance](image)

Figure 2.11 $I$-$V$ of a device showing negative differential resistance [105].

**Transferred electron devices (Gunn diodes)**

The Gunn diode, named after J. B. Gunn [96] who discovered the effect, is an active solid-state two terminal device, classed as a transferred electron device (TED). The most distinct feature of the device is the negative differential resistance, which depends on bulk material properties [96]. The physics of the Gunn effect will be elaborated in 2.3.

The Gunn diode in its basic form is a homogenous two terminal device with an ohmic contact at each end. The device has a sandwich-like structure and comprises of $n^+\cdot n\cdot n^+$ semiconductor materials and is usually grown using Molecular Beam
2.2 THz generation

Epitaxy (MBE). GaAs and InP are the most commonly used material systems although other materials such as CdTe, ZnSe, GaAsP and GaN may possibly also exhibit the transferred electron effect and are being researched for use at higher frequencies [106]. State-of-the-art GaAs and InP Gunn diode output powers, and DC to RF conversion efficiencies are shown in Fig. 2.12. The output power falls off as $1/f^2$ and then as $1/f^3$ as the frequency increases due to material limitations such as energy relaxation time. Frequency multipliers with two or more diodes are generally employed to reach frequencies above 0.2 THz, up to about 1 THz. The average power level achievable in the region around 0.400 THz is typically in the range 0.1 to 1 mW [107].

![Figure 2.12](image)

**Figure 2.12** State of the art results from GaAs and InP Gunn devices. Numbers next to the symbols denote DC to RF conversion efficiencies in percent [104].

Though simple in physical structure, GaAs or InP Gunn diode devices require infinite care during growth for optimisation of physical parameters such as doping concentration and device length. The ohmic contact resistance needs to be as small as possible, as this can considerably affect device resistance and consequently the operating frequency range. Devices operating in W-band (75 to 110 GHz) typically have specific contact resistivity values of less than $5 \times 10^{-6} \, \Omega \text{cm}^{-2}$, while for D-band and above (i.e. above 110 GHz), $5 \times 10^{-7} \, \Omega \text{cm}^{-2}$ or less is needed [108].
Currently, commercial GaAs Gunn diodes provide moderate power levels of ~60 mW at 94 GHz in second harmonic mode [109]. Due to low phase noise (-88 dBc/Hz at 100 KHz offset) at mm-wave frequencies, Gunn diodes are considered extremely suitable for Frequency Modulated Continuous Wave (FMCW) radar and imaging systems. However, the major issue of the Gunn diode is its low DC-to-RF conversion efficiency and associated high operational temperatures. In response to high costs associated with fabrication, physical modelling of devices offers an alternative to predict behaviour and performance before experimentally verifying them.

**Tunnelling devices**

Both Esaki tunnel diodes (TDs) [97] and resonant tunnelling diodes (RTDs) [98] are categorised as Tunnelling devices and work on different tunnelling mechanisms. The TD comprises of a heavily doped (degenerate) p⁺-n⁻ junction where interband tunnelling takes place from the valence band to the conduction band. In a RTD, the tunnelling occurs in the conduction bands of a double-barrier heterostructure.

Compared to other solid-state two terminal devices, both TDs and RTDs provide very low output power. The output power is limited by the small voltage swing (RF) and large junction capacitance. State-of-the-art RTDs have achieved power levels of 0.3 µW at 712 GHz using InAs/AISb RTD [110] and 1 µW at 831 GHz using GalnAs/AlAs double-barrier RTD [111].

**Transit-time diodes**

Devices that use special current injection mechanism fall under the transit-time diode category. These devices have carriers injected into a depletion region which drift through the device active region with the drift velocity. The drift velocity depends on the applied electric field in the active region. This creates a phase shift between device terminal voltage and current, which in turn creates a NDR and generates RF oscillations [20].

Various transit-time diodes carrier generation and injection are listed below:

- In IMPATT diodes [99] carriers are formed and injected due to avalanche multiplication through impact ionisation occurring in a reverse-biased p-n junction.
2.2 THz generation

These devices are acknowledged to be capable of providing high power at mm-wave frequencies. However, due to the avalanche effect, they not only require a high current source to operate but also have an inherently high phase-noise. Thus, their application as Local Oscillators (LOs) in FMCW systems is not a viable option. However, presently Si IMPATT diodes are being used in passive mm-wave radiometric imaging systems as incoherent noise sources for illumination [112].

- The BARITT diodes [100] generate microwave oscillations by using thermionic emission of carriers over a forward biased barrier. The barrier is formed due to p-n or Schottky junction or heterojunction, which contributes to positive active resistance by forming an RC circuit. Other BARITT diode limitations include longer active region, small NDR, low output power and efficiency as compared to IMPATT diodes [20].

- The TUNNETT diodes [101] have carriers injected by tunnelling. The tunnelling is band-to-band in case of a p-n junction and through the barrier for Schottky barriers. The TUNNETT diodes have lower noise than IMPATT diodes but are limited by low output power due to low tunnelling current. However, they can work at lower operating voltages and theoretically can achieve 1 THz. Experimentally, it has been shown that the TUNNETT diodes have achieved power levels of 7.9 mW at 655 GHz [113].

- MITTATT diodes [103, 113] use both tunnelling and impact ionisation mechanism for carrier generation. These devices have smaller carrier generation region and are designed for high frequency operation [102]. Their limitations include high noise level, small NDR, series resistance and poor impedance matching characteristics. It has been shown that the MITATT diodes have achieved power levels of 3 mW at 150 GHz [114].

Monolithic Microwave Integrated Circuit (MMIC)

MMIC solutions are typically chosen over hybrid technologies due to their reliability and increased functionality. This was reflected in the 2004 GaAs device market segmentation, which showed that 83% of the market was taken by MMIC technology, with the remaining 16% and 1% accounted for by discrete and digital ICs respectively [66].
2.2 THz generation

As opposed to Gunn diode based systems, MMIC solutions have a compact design, can be easily integrated into planar circuits, do not require a cavity and can provide a higher degree of functionality (Fig. 2.13). However, commercial MMIC solutions have yet to reach the output power levels provided by Gunn diodes at mm-wave frequencies [60] and generally cost more. In recent times, planar Gunn diodes work has produced encouraging results [115-118] though the output power levels are still low. The output power of a state-of-the-art planar Gunn diode has been reported as 1.58 µW at 115.5GHz [119].

Commercially available low phase-noise pHEMT-based MMIC chip-sets generating a frequency of 77 GHz have been developed for automotive and general radar applications, where they have slightly lower phase noise than Gunn diode based oscillators. United Monolithic Semiconductors (UMS) has developed a MMIC based FMCW chipset using a low frequency oscillator MMIC (38.5 GHz) and a two times frequency multiplier MMIC [66].

![Figure 2.13](image)

Figure 2.13  (a) 3D view of a MMIC power amplifier in a package [120]. (b) Scanning Electron Microscope image of a planar GaAs-based MMIC working at 0.7THz to be integrated with amplifiers under the DARPA Terahertz Transistor and Imaging Program. This represents the first wafer-bonded all-planar mixers and amplifiers for a 0.7 THz receiver system [120].

The emission powers of the various THz sources discussed in this chapter are illustrated as a function of frequency in Fig. 2.14. The solid lines and ovals denote conventional and more recent THz sources respectively.
2.3 The Gunn effect

The voltage oscillation that occurs when an applied bias voltage exceeds a certain threshold, is called the Gunn effect, named after its discoverer J. B. Gunn [96]. The free-running frequency of the oscillation depends on the material properties and the geometry of the device itself. This is due to the transferred electron effect exhibited by certain binary and ternary compound semiconductors. A solid-state two-terminal device that converts DC to RF using this effect is commonly referred to as a Gunn diode. The transferred electron effect principle is presented here discussing the low- and high-field electron transport, negative differential mobility and instability leading to domain formation in Gunn diodes.

2.3.1 Physics

Low-field electron transport and drift velocity

At low electric fields, the electron drift velocity is linearly proportional to the applied electric field. However, it departs from the linear relationship with the
2.3 The Gunn effect

increase in applied electric field, which causes the drift velocity to approach the thermal velocity value. The low-field electron drift velocity is given as [112]:

\[ v_D = -\left( \frac{q \tau}{m^*} \right) E \]  \hspace{1cm} (2.2)

where \( v_D \) is electron drift velocity, \( E \) is the applied electric field, \( \tau \) is the mean free time and \( m^* \) is the effective electron mass. The drift velocity is proportional to the electric field and the proportionality factor is known as the mobility, which is given by:

\[ \mu = \frac{q \tau}{m^*} \]  \hspace{1cm} (2.3)

Thus the drift velocity can be written as:

\[ v = -\mu E \]  \hspace{1cm} (2.4)

where the negative sign indicates that the electrons move in the opposite direction to the applied electric field. The mobility determines the influence of electric field on the motion of an electron. At low electric field values the drift velocity of electrons is linearly proportional to the applied electric field and is smaller than the thermal carrier velocity (which is \( 7.7 \times 10^6 \) cms\(^{-1} \)) for n-type Gallium Arsenide (GaAs) [121]. As the electron drift velocity approaches the thermal velocity, the linear relationship due to the corresponding constant mobility breaks down. The drift velocity increases with the increase in electric field until reaching a saturation velocity value. The plots of carrier drift velocity versus electric field for GaAs, Indium Phosphide (InP) and Gallium Nitride (GaN) are shown in Fig. 2.15, where it can be seen that the drift velocity after reaching a maximum value decreases with an increase in applied field.

**High-field electron transport**

The negative differential mobility phenomenon can be explained using band diagrams of the different III-V semiconductors shown in Fig. 2.16. Fig. 2.16a shows that GaAs
2.3 The Gunn effect

Figure 2.15 Carrier drift velocity vs electric field for three semiconductor materials: GaAs, InP and GaN [121]

![Carrier drift velocity vs electric field](image)

**Figure 2.16** Electron occupations for (a) GaAs, (b) InP and (c) GaN [112]

![Electron occupations](image)

is a direct band gap material with an energy gap, between the top of valence band and bottom of conduction band, of 1.43 eV at \( k = 0 \) (zero wave vector). The satellite conduction band minimum (L) is 0.32 eV higher than the conduction band minimum (Γ) and is at \( k = \pm \frac{2\pi}{a} \) (where \( a \) is the lattice constant). At 300 K, n-type GaAs has its valence band highly occupied with few electrons in the conduction minimum.
(central Γ band) and nil in the satellite L band. However, as the increase in applied electric field reaches ~3 kV/cm, some electrons from the Γ valley gain sufficient energy and are scattered into the L valley. As shown by Eq. 2.3, electron mobility is dependent on effective electron mass, which is different in the different conduction band valleys and depends on the local curvature of the band structure. The effective electron mass is given by [106]:

\[ m^* = \frac{1}{2\hbar^2} \left( \frac{\partial^2 E}{\partial k^2} \right) \]

where \( m \) is the effective mass, \( \hbar = h / 2\pi \) is plank’s constant and \( k \) is the wave vector. In the satellite valley the curvature is higher than the central valley, which results in a higher effective mass compared to the central Γ valley. For GaAs, electrons have an effective mass of \( 0.4m_0 (m_1^*) \) in the L valley and \( 0.068m_0 (m_2^*) \) in the Γ valley [122].

The density of states of each Γ and L valley represent a number which is proportional to energy interval available to the conduction electrons and number of allowed energy states per unit volume. Its ratio, \( R \), is given by:

\[ R = \frac{M_2}{M_1} \left( \frac{m_2^*}{m_1^*} \right)^{\frac{3}{2}} \]

where \( M_1 \) and \( M_2 \) are the equivalent Γ and L valley whose values for GaAs are 1 and 4 respectively [20]. Thus the density of states ratio becomes \( R = 94 \) [20], which is much higher than the corresponding Γ valley. In the L valley in addition to the higher electron effective mass, the electrons are also subjected to stronger scattering processes [123]. These effects collectively result in lower electron mobility (up to 70 times) in the L valley compared to the Γ valley.

Negative differential mobility

In Fig. 2.17a, applied electric-field is low \( (E < E_\alpha) \) thus all the electrons are in the central valley. As the field is increased \( (E_\alpha < E < E_\theta) \), some electrons gain the energy required to transfer to upper valley as shown in Fig. 2.17b. Finally when the
2.3 The Gunn effect

Field is large enough \((E > E_\beta)\), effectively all the electrons are transferred to the upper valley. The effect of this on the drift velocity is shown in Fig. 2.18, which can be equated as [112]:

\[
\nu_d \approx u_1 E \quad \text{when } 0 < E < E_\alpha
\]  

\[
\nu_d \approx u_2 E \quad \text{when } E > E_\beta
\]  

The electron drift velocity increases linearly until the onset of negative differential mobility where the electric field is \(E_\alpha\). The drift velocity continues to decrease until the field acquires the maximum value \(E_\beta\) as shown in Fig. 2.18. It can be seen that \(u_1 E_\alpha\) is greater than \(u_2 E_\beta\), which results in a decrease of drift velocity between \(E_\alpha\) and \(E_\beta\). The negative differential mobility exists between the threshold field \((E_T)\) and final valley field \((E_V)\) [112, 121].

![Figure 2.17](image)

**Figure 2.17** Electron occupations under various electric field levels for n-GaAs [112].
2.3 The Gunn effect

![Diagram showing negative differential mobility region](image)

**Figure 2.18** Negative differential mobility region [112].

If the electron mobility, effective mass and density for both central (\( \Gamma \) band) and satellite (\( L \) band) valleys are denoted by \( \mu_1, m_1, n_1 \) and \( \mu_2, m_2, n_2 \) respectively then the steady state conductivity of n-type GaAs is given as [112]:

\[
\sigma = q(\mu_1 n_1 + \mu_2 n_2) = qn\bar{\mu}
\]  

where \( \bar{\mu} \) is the average mobility and is given by:

\[
\bar{\mu} = \frac{(\mu_1 n_1 + \mu_2 n_2)}{(n_1 + n_2)}
\]  

Thus the average drift velocity can be written as:

\[
\nu_d = \bar{\mu}E
\]
2.3 The Gunn effect

Conditions for negative differential mobility

For a material to exhibit the transferred electron effect and therefore negative differential mobility, it must display the following characteristics [124]: It must have at least two valleys in the conduction band.

i. The upper valley ($\Gamma$) must be at least several $kT$ above the satellite ($L$) valley in the conduction band at the lattice temperature $T$. This is to avoid inter-valley transfer due to thermal occupations at low electric field, ensuring it occurs only due to the applied electric-field.

ii. The energy difference between the $\Gamma$ and $L$ valley must be less than the bandgap energy $E_g$ in order to avoid impact ionisation of electrons across $E_g$ before inter-valley transfer occurs.

iii. The inter-valley transfer must take place in a time that is much less than the period of the operating frequency.

Other than GaAs, InP and GaN depict the above characteristics and thus binary and ternary compounds made of these materials are often used in transferred electron devices [112, 121, 125].

Instability and domain formation

Transferred electron devices are inherently unstable. The instability is due to the formation of a space charge region that grows exponentially with time due to a small, random change in carrier density in the semiconductor. This instability as explained in Ref. 122 is discussed below.

In a uniformly doped semiconductor device with an electron concentration $N_D$, a noise process or variation of doping level or some crystal defects can cause a spontaneous fluctuation in the electron density ($N$). This spontaneous fluctuation forms a dipole, consisting of an accumulation and a depletion region with associated electric field gradient. The non-uniformity in the space charge is related to the electric field by the following Poisson equation:
2.3 The Gunn effect

\[ q(N - N_D) = E_0 E_r \frac{\partial E}{\partial x} \]  

(2.12)

where \( E_0 \) and \( E_r = \) are the vacuum and relative permittivity respectively, and \( N_D \) is the doping density. If the mean electric field is less than the threshold field then electrons subjected to a higher electric field will move faster than those in the low-field region. The space charge would then fill the depletion region resulting in the dampening of the fluctuation due to dielectric relaxation.

Now considering the case when the applied electric field is greater than the threshold electric field \( E_T \): the electron drift velocity reduces in the region of higher field, causing the space charge fluctuation to grow along with time associated local electric field in the region as shown in Fig. 2.19b. This exponential growth of electric field continues until a stable domain is formed as shown in Fig. 2.19c. To keep the total voltage drop across the device the same, the electric field outside the domain remains below the threshold level, hence achieving stability. It can be seen in Fig. 2.19c that the electric field within the domain reaches a peak value \( E_p \), while outside the domain, the electric field \( (E_R) \) is less than the threshold value. The space charge present due to the velocity difference is stabilised by the accommodation in the electric field \( (E_R < E_T) \) and a stable domain is formed. Additionally this means that when the domain is stable another domain does not grow as \( E_R \) remains below \( E_T \).

Figure 2.19 Stable dipole domain formation with the growth of space charge [122].
2.3 The Gunn effect

Generally in GaAs, $v_D \approx 1 \times 10^7 \text{ cms}^{-1}$ for moderately doped material. As the domain grows, the current in an external circuit falls due to the fall in the field outside the domain. As the domain reaches the anode, the electric field gradient becomes zero and the domain collapses. The external electric field and current values increase until reaching the threshold value and allowing another domain to form at the cathode. The external current therefore oscillates at a frequency, which is inversely proportional to the transit length $l$, (the distance between the domain formation point and anode). The oscillating frequency can therefore be given as:

$$f_0 = \frac{v_D}{l_t}$$  \hspace{1cm} (2.13)

2.3.2 Conventional Gunn diode and limitations

The Gunn diode in its basic form is a homogenous device with ohmic contacts at each end. The device has a sandwiched structure and comprises of $n^+$ contact layers at the ends with a $n^-$ transit region in between, as shown in Fig. 2.20. As previously described, the transit region length partially determines the device’s operating frequency. However, domain formation in conventional Gunn diodes does not occur in the initial portion of the transit region, where the electrons are being accelerated to the energy levels required for inter-valley transfer. This initial portion is termed as the dead zone. Thus domain formation occurs at a certain distance away from the cathode depending on the applied bias. This reduces the effective length of the transit region and acts as parasitic positive resistance, which degrades the overall negative resistance and results in a reduction in DC-to-RF conversion efficiency. Generally the length of the dead zone is about 40 percent of the transit length in high frequency devices depending on the transit length and the applied bias [126]. The applied bias determines the point of domain nucleation, making the effective transit length and therefore operating frequency very sensitive to the applied bias.
2.3 The Gunn effect

Conventional GaAs Gunn diodes can provide around 500 mW power at 20 GHz with 5.5\% efficiency. The efficiency reduces to about 1.5 \% and the power to 50 mW at 60 GHz. It has also been observed that at 77 GHz with 0.8 \% efficiency, powers of up to 50 mW have been achieved. The conventional Gunn diode limitations are as follows [127]:

- High turn on voltage.
- Poor bias and temperature stability
- High thermal parasitic value at high frequencies.
- High operating current leading to more power dissipation.
- Low efficiency and power derived at millimetre-wave frequencies.

2.3.3 Planar Gunn devices

The initial interest in developing planar Gunn devices stemmed from the potential application for high-speed logic devices as the planar geometry was ideal for mass production. However, the research on these devices for logic circuits was hindered in the 1970s by semiconductor material and device contact issues. But in recent times, they have regained the attention of researchers due to the high demand for millimetre-wave and THz sources. In planar Gunn devices, the current flow is parallel to the epitaxial layers as shown in Fig. 2.21; in contrast to the perpendicular flow in vertical devices. Planar devices meet the high frequency generation requirements and show advantages over vertical devices; especially with advances in
2.3 The Gunn effect

wafer growth and fabrication technology, nanoscale devices using high quality semiconductor materials are easily and reliably achievable.

Figure 2.21 Schematic of electron flow in (a) a vertical Gunn diode and (b) a planar Gunn diode.

Compared with conventional vertical Gunn devices, planar Gunn diodes have several advantages. Firstly, they have lithographically controlled anode-cathode separation that determines the oscillation frequency of the device. This is not the case with the conventional Gunn diode, because once the wafer is grown, the anode-cathode separation is fixed. In other words, the transit-time oscillation frequency is fixed unless appropriate tuning circuits are employed; even then the circuit tuning is very limited. On the other hand, the planar structures allow great flexibility in adjusting the anode and cathode distance and therefore the oscillation frequencies. Potentially, such devices may oscillate at several hundreds of gigahertz or even THz frequencies once the distance is further reduced to submicron dimensions. Secondly, the planar structures are compatible with other planar circuitries, such as coplanar waveguide (CPW)-based components, so that complex circuits and systems, such as transceivers, can be fabricated on a single chip without integration issues. Such seamless connection between signal sources and monolithic microwave integrated circuits (MMICs) will significantly improve the productivity and reproducibility, which is not achievable with conventional vertical Gunn devices because each individual device has to be cleaved and encapsulated in a cavity. Certainly, the planar Gunn diodes may face challenges, such as low power or low phase noise...
2.4 Self-switching device (SSD)

compared to conventional Gunn diodes. However, planar structures allow a combination of a large volume of devices to improve the power performance.

The concept of a self-switching device (SSD) was introduced in 2003 by A.M. Song et al [46, 47]. It is a nonlinear nanodevice where its $I-V$ characteristic (Fig. 2.22a) follows that of a conventional diode but without using a doping junction or a tunnelling barrier. Moreover, the turn-on voltage can be tuned by simply altering the width of the channel.

![Figure 2.22](image)

**Figure 2.22** (a) A typical $I-V$ characteristic of the self-switching diode (SSD). The SSD was fabricated on InGaAs/InP substrate with a 1.2 μm channel length ($L$) and 80 nm channel width ($W$). The measurement was performed at a low temperature of 4.2 K. (b) Illustration of the depletion regions at the boundaries caused by the charges at the surface states of the etched trenches. Depending on the sign of the applied voltage $V$, the effective channel width will (c) reduce or (d) increase resulting in the diode-like characteristic [46].
2.4 Self-switching device (SSD)

Figure 2.23 I-V characteristics of two SSDs with different $W$ and $L = 1.2 \, \mu m$. Both SSDs were fabricated on InGaAs/InAlAs substrate. Turn-on voltage decreases as $W$ increases from 60 nm to 70 nm [46].

Since its inception, the properties and performance of the device has been explored in numerous experiments [46-49, 51, 52, 128-141]; and extensive characterisation studies performed using modeling [128, 129, 142] and simulations [50, 133, 142-165]. More recent developments include room temperature detection at microwave [129, 130] and THz [49, 137] frequencies; and Monte Carlo (MC) simulations [50, 155, 161, 162] showing THz frequency oscillations in the device making it a potential emitter in this range.

The SSD is a unipolar two-terminal device as shown in Fig. 2.22b. Two L-shaped insulating trenches are etched through the semiconductor forming a non-symmetrical layer with a defined conducting channel between the terminals. The effective channel width of the SSD is actually narrower than its physical size, because of the depletion regions at the etched boundaries. These depletion regions are caused by the charges at the surface states of the insulating trenches. As can be seen in Fig. 2.22a, the effective channel is almost pinched-off at unbiased condition due to these depletion regions. A negative bias applied to the anode induces negative charges around the trenches thus depleting the channel and impeding the current flow, Fig. 2.22c. In contrast, a positive bias counteracts the lateral depletion, widening the effective width of the channel and allowing current to flow easily, Fig. 2.22d. This self-switching mechanism results in a strongly nonlinear current-voltage (I-V) characteristic. Unlike a conventional diode, the working principle of the SSD is based on geometrical symmetry breaking at nanoscale. The SSD can operate with a geometrically defined zero-voltage
2.4 Self-switching device (SSD)

threshold independent of the material used (Fig. 2.23), making it ideal for microwave applications.

SSDs have been demonstrated in a variety of materials including two-dimensional electron gas (2DEG) in GaAs [48], InGaAs [46, 130] and GaN [137], silicon-on-insulator (SOI) [128, 133], and both organic [51, 135] and metal oxide [136, 138] thin films. The active area of the SSD is easily defined by a single high-resolution lithography step, such as electron beam lithography, Fig. 2.24a. The single-step process simplifies and makes the manufacturing of the nanodiodes compatible with cost-effective fabrication methods, such as nanoimprint lithography (NIL) [51, 166]. A large array of SSDs can be fabricated in parallel with one another in a single step without the need for interconnection [130, 134] which reduces the impedance of the overall device favorably, Fig. 2.24b. The device geometry can also benefit from multilevel NIL [167, 168], which can be potentially utilised to define high density integrated structures at a relatively low cost.

![Figure 2.24](image)

Figure 2.24 (a) Atomic-force micrograph of two SSDs connected in parallel fabricated using electron beam lithography. The channel length and width are 1.5 μm and 130 nm, respectively [48]. (b) SEM image of 100 SSDs connected in parallel, fabricated using a single lithography step without interconnection layers [131].

2.4.1 Applications

The nonlinear behaviour of the SSD provides perhaps the most fundamental application of functioning as a rectifier in a lot of the work carried out so far [48, 128, 130, 135, 137, 138]. However, the greatest advantage over conventional
rectifiers is its planar architecture, providing intrinsically low parasitic capacitance, thus allowing rectification at very high speeds (1.5 THz at room temperature [48] and 2.5 THz below 150 K [134]), even with low mobility materials (>50 MHz with ZnO [138]). Significantly more applications can be realised by simply manipulating the geometry of the device. SSD properties related to the length and width of the channel; and the width and dielectric constant of the insulating trenches, etc. have been extensively studied using MC simulations [144, 152]. Most recently, a zero-turn-on voltage attained by carefully defining the channel width was used for radio-frequency (RF) rectification without requiring any external power [169].

Since the carrier transport mechanism in the SSD is solely dependent on surface charges and electric field effects and does not require ballistic transport [46, 133], existing technology such as complementary-metal-oxide-semiconductors (CMOS) can be harnessed quickly for use with SSDs. This has already been demonstrated with SOI-based SSDs as logic circuits [128], as well as RF detectors with a cut-off frequency of 5 GHz [129]. Further simulation work on SOI-based SSDs carried out by Farhi et al, using Taurus Medici from Synopsys, can be found in Refs. [132, 133, 164].

Interestingly, SSDs designed on low-cost semiconductor materials such as polymer and metal-oxide (ZnO) thin films have demonstrated RF rectification, even though their mobilities are relatively low [135, 138]. This will open up opportunities in the RF fields where simple and low cost technology is required. An ideal application would be to use them for RF identification tags which have received much attention [170, 171]. Additionally, the fabrication of more SSDs into an array will provide better noise equivalent power (NEP) and signal to noise ratio (SNR), as well as improve the impedance matching [134]. Recent low-frequency measurements using an SSD array showed lower noise spectra and NEP comparable to state-of-the-art Schottky diodes [49]. The result has great implications for high frequency applications, such as in the THz region, where the novelty of the planar architecture will be a tremendous advantage compared to many conventional solid-state two-terminal active devices. More work on low- and high-frequency noise properties in SSDs performed using MC simulations can be found in Refs. [147, 150, 156-160]. As mentioned earlier, high speed rectification at THz frequencies...
2.4 Self-switching device (SSD)

has already proved the SSD’s potential to be used as a detector. These devices comprise of high mobility 2DEG in InGaAs/InP[134] and AlGaAs/GaAs [48]. Using these and other III-V materials, which exhibit negative differential mobility, MC simulations have shown that SSDs can be used as THz emitters via Gunn oscillations [50, 153, 155, 162].

If successful, the SSD can fill the gap of the much sought after low-cost, compact THz emitters; and possibly achieve high efficiency owing to the numerous advantages listed above. Most importantly, being a room temperature solid-state device will attract avid interest from all of the applications discussed in 2.1, and potential new ones as well. So far, however, no systematic studies on the device geometrical parameters and especially the influence of interface-charge density have been reported, which is one of the objectives of this work. The study aims to identify geometry dependencies of the Gunn oscillations in the SSDs so that, by constructing an array that contains different geometries of SSDs placed in parallel, frequency tuning in wide or narrow bands can become possible, which may have useful implications to practical applications. The surface charges in the channel boundaries play such a major role in the electron transport of the device that its effects cannot be neglected. Hence, the characterisation of emission in different space charge environments is considered very crucial to the overall analysis.

Last but not least, the SSD structure can also be used as a memory device by utilising a hysteresis effect that occurs in the SSD under certain conditions [52]. This property of the SSD is the motivation behind further developments carried out during this research work to harness the memory effect for potential applications. A discussion of the previous results for Gunn oscillations and the memory effect obtained from SSDs is given in the next sections.

2.4.2 Gunn oscillations

MC simulations have shown that Gunn oscillations can occur in SSDs under appropriate conditions, the frequencies of which are in the RF and very low THz range. The implications are significant considering the need for solid-state THz frequency emitters is paramount for the present and the future. The oscillations derived from the SSD in MC studies conducted independently by K.Y. Xu et al [50]
2.4 Self-switching device (SSD)

and T. González et al [155] are shown in Figs. 2.25 and 2.26 respectively. Both work was performed using a similar device structure based on InGaAs of channel length ~1.2 μm and channel width 60 nm; hence achieving oscillations that had identical characteristics of ~0.1 THz above a bias of 2.5 V. However, in Ref. 155, a frequency of 0.4 THz was demonstrated in a GaN based SSD owing to its wider bandgap, higher saturation velocity and shorter energy relaxation time. The study by K.Y. Xu also reported on the factors that may be responsible for generation of the Gunn oscillations. Firstly, the asymmetric boundary conditions of the SSD result in a strong nonuniform distribution of electric field along the nanochannel [148], which is a favorable condition for Gunn oscillation. The effective electron channel in the SSD is also widened by a positive bias as opposed to a symmetric device. This particularly occurs around the left entrance of the nanochannel due to the strongest transverse electric field, which allows more electrons to flood into the nanochannel to form charge domains.

Figure 2.25 (a) Schematic top view (not to scale) of the simulated SSD. (b) Current responses of a SSD compared to a symmetric structure. The applied voltage changes from 2.5 to 3.0 V at time zero [50].

Figure 2.26 (a) Geometry of the InGaAs SSD. (b) Current sequence for ΔV = 0.5 V applied every 50 ps [155].
2.4 Self-switching device (SSD)

The preliminary investigations lead to foreseeable improvements that can be made to the device architecture to gain more efficient performances out of the planar SSD. For one, a large array of SSDs can be fabricated in parallel with one another in a single step of nanolithography without the need for interconnection [130, 134]. In such a case, the impedance of the overall device will be favourably reduced. More importantly, the output power can be enhanced by a factor at least equal to the number of SSDs or even more if there is any phase locking of the oscillations in individual nanochannels. It is also possible to make the individual SSDs with different nanochannel widths so as to generate different oscillation frequencies on a single chip. The planar architecture also allows radiation naturally along the normal direction of the device surface due to the in-plane oscillating electric field. This allows for coupling of a suitable antenna to the device to assist the radiation. In these respects, benefits such as above are difficult to achieve in conventional vertical Gunn diode structures since the oscillation frequency is generally fixed by a given semiconductor wafer.

Therefore, it was considered an essential step, in the progress of defining the characteristics of Gunn oscillations in SSDs, to report how geometrical parameters affect different performance factors such as oscillation frequency, bias requirements, etc. In this work, the two main dimensional parameters of channel length and width are varied and their performances recorded using SILVACO, Atlas, a much approved simulation tool for semiconductor technology at industry level. The critical interface-charge density along interface boundaries also largely determines the electrical properties and electron transport in the device [46, 133]; hence the appraisal of their presence in influencing oscillation properties is also duly performed.

2.4.3 Memory effect

The SSD structure can also be used as a memory device by utilising a hysteretic effect that occurs in the SSD under certain conditions, as shown in Fig. 2.27 [52]. A pronounced hysteresis begins to emerge when the applied negative voltage has passed beyond the current breakdown at about - 0.9 V. Normally, the SSD $I-V$ curve (i.e., the dashed line) is independent of the voltage sweep direction. However, once
2.4 Self-switching device (SSD)

the breakdown has occurred, sweeping the applied voltage in the opposite direction yields a very different $I$-$V$ characteristic, as indicated by the solid line in Fig. 2.27.

![Figure 2.27](image)

**Figure 2.27** Hysteresis effect in the $I$-$V$ characteristic of a typical self-switching diode (SSD) that can be utilised for memory operations. Both dashed and solid arrows indicate the sweeping direction of the applied voltage. The measurement was carried out at temperature of 24 K [52].

By assigning states of 0 and 1, based on the current levels in the I-V curve, for example at $-0.5$ V, memory operations using the SSD can be realized. Memory retention times recorded were in the order of minutes at room temperature and hours at cryogenic conditions. The hysteresis was proposed to be a result of charging and discharging of surface states along the nanochannel also verified by means of MC simulations[146, 149]. Such memory devices consisting of just two terminals, also termed memristors, though seemingly sufficient, suffer from drawbacks in memory applications. This is because only one terminal can be used for both reading and writing, thus posing difficulties incorporating them into a large network of memory cells, and would require new memory architecture in circuitry. Hence, a three terminal extension of the SSD would be more suitable yet preserving the planar nature of the device. For this, a novel device is conceptualised and studied in detail for memory applications, Fig 2.28.
The device is generically based on the SSD but introduces a third terminal to be used as a control gate for the current in the channel. The novelty in this concept is the presence of a memory storage region adjacent to the nanochannel functioning as a floating gate instead of an oxide layer conventionally used in flash memory technologies. Self-terminating encapsulation of charge is made possible with the presence of two nanogaps around the perimeter of the storage region. Owing to their dimensions and surface states present on the walls of the nanogaps, charging and discharging of the surface states can alter the effective gap sizes (creating barriers of Detailed characterisation studies of the device are performed and memory effects are investigated in terms of memory retention times, control gate pulses width and bias voltages.
3
Development of ATLAS for planar nanodevices

3.1 Introduction

Computation/simulation using modelling enables quick design, test, and evaluation of semiconductor devices without physically fabricating them. This translates into huge advantages in terms of cost reduction and vastly reduced time to market, as different device structures can be studied in a very short time. Simulation packages available now are very powerful tools that meet industrial standards. These are often used to obtain device performance parameters such as potential, carrier and current profiles within the device during operation, which cannot be measured experimentally.

Semiconductor devices can be modelled in two ways; one is to determine the terminal electrical properties of a device based on empirical results and model the device using techniques like spline fitting. The other method is to study the carrier transport processes taking place within the device. The latter, which is physics-based and employed in this work, provides three major advantages: 1) it is predictive, 2) it provides insight, and 3) it captures knowledge in a way that makes it easier to understand the various phenomena acting within the device.

1 Spline fitting is a method of fitting a smooth curve to a set of noisy observation using a spline function – a smooth piecewise polynomial that can be used to represent functions over large intervals, where it would be impractical to use a single approximating polynomial.
3.2 SILVACO

The simulation software package used in this work is a commercial semiconductor device simulator ATLAS, from SILVACO, Inc. SILVACO packages have become a major tool for semiconductor companies to acquire physically-based simulation capabilities as semiconductor devices are becoming smaller and faster; and integrated circuit design and performance become significantly more reliant on process related effects. SILVACO is also the leading vendor among Technology Computer Aided Design (TCAD) software for semiconductor/electronics simulations. Some of the major tools included in this software are ATHENA, ATLAS, DECKBUILD, TONYPLOT, etc. Throughout this work, the devices were built using DECKBUILD and the simulations were run in ATLAS. The outputs were viewed using TONYPLOT. A brief on the tools employed is given next.

DECKBUILD

The Interactive Runtime Environment is the central environment for interactively using process and device simulators. It provides many important capabilities. A graphical user interface for input deck specification allows users to avoid simulator-specific input syntax. Information is typed into a series of pop-up windows. When specification is complete, DECKBUILD automatically produces a syntactically correct input deck. Decks can be edited by the user at any time. Multiple decks are produced if input parameters are looped, and parameters can be extracted from the calculated results. Multiple simulators can be called from within a single input deck, and information transfer between simulators is transparent to the user. DECKBUILD allows precise user control of how an input deck is run, stopped, paused or restarted within a single step. It provides a history function that permits the user to backtrack to a previous point in the deck, and then continue computation from this previous point. This capability is extremely useful for interactively developing a simulated process flow specification.
SILVACO’s ATLAS is a semiconductor simulator that has been used extensively in academia and industry. Notable companies that use ATLAS are Advanced Micro Devices Corporation, Harris Semiconductor, NASA, Sensitron Semiconductor, and Science Applications International Corporation (SAIC). Academic institutions that use ATLAS include the University of California at Berkeley, Stanford University, and the University of California at San Diego. ATLAS has been continually refined since the 1970’s to be more realistic for a wide range of semiconductor device configurations.

ATLAS is a versatile, modular and extendable software product for two and three dimensional device simulation. Simulation is done by discretising the equations that describe the operation of a device onto a two or three-dimensional grid, consisting of a large number grid points called nodes. The ATLAS simulation package is built around physics-based models. These physics-based models are basically a mathematical description of the various phenomena, such as carrier generation, recombination and transport that are responsible for the operation of the device. The models describe the dynamics of carrier generation, recombination and injection as well as transport within different material regions of the device and coupling between the regions. These models are quantified in differential equations and transposed onto the grid to simulate the transport of carriers through a device structure. Electron and hole concentrations are calculated, along with their spatial variation and dependence on material properties as well as biasing. The simulator also predicts the electrical characteristics of the device, for example the terminal currents that are associated with the specified physical structures and different bias conditions. These results can be used to characterize the DC and AC operation of the device and to provide insights into the physical aspects of device operation, provided they include all the important device physics. These models consist of a set of fundamental equations, which link together the electrostatic potential, charge density and the carrier densities within a simulation domain. An important advantage of this type of model is that it can not only be used to predict the terminal current-voltage

---

^2Discretisation concerns the process of transferring continuous models and equations into discrete counterparts. This process is usually carried out as a first step toward making them suitable for numerical evaluation and implementation on digital computers.
3.2 SILVACO

characteristics of the devices for comparison with device measurements, but also to
investigate the origins of the observed behaviour and the factors limiting the device
performance.

ATLAS solves three fundamental types of equations when simulating a
semiconductor device: Poisson’s equation, the continuity equations, and the transport
equations, including the Drift-Diffusion and Energy Balance Transport Models
which will be explained later in this chapter. Poisson’s equation relates the space
charge density to electrostatic potential, and allows calculation of electric fields
based on the potential. The transport equations are all derivative of the Boltzmann
transport equation. For most applications, the Drift-Diffusion Transport Model is
used. Only for small, submicron devices are the Energy Balance or Hydrodynamic
Models considered necessary for the simulations.

ATLAS inputs and outputs

As shown below, most ATLAS simulations use two inputs: a text file that contains
commands for ATLAS to execute, and a structure file that defines the structure that
will be simulated. ATLAS produces three types of output. The run-time output
provides a guide to the progress of simulations running, and is where error messages
and warning messages appear. Log files store all terminal voltages and currents from
the device analysis, and the solution files store two and three-dimensional data
relating to the values of solution variables within the device for a single bias point.

Figure 3.1 ATLAS inputs and outputs [172].
3.3 Device physics

Each input file must contain these five groups in order. Failure to do this will usually cause an error message and termination of the program, but it could lead to incorrect operation of the program. For example, material parameters or models entered in the wrong order may not be used in the calculations. The order of statements within the mesh definition, structural definition, and solution groups is also important. ATLAS uses its sub-engine ‘Blaze’ for heterostructure device simulations. Blaze is a general purpose 2D device simulator used for III-V materials and is invoked by default for heterostructures with a position dependent band structure. During simulations it modifies the charge transport equations to account for the effects of positional dependent band structures [172].

TONYPLOT

TONYPLOT (visualization tool) provides comprehensive interactive scientific visualisation capabilities. All of the usual ways of displaying scientific data are supported. These include x-y plots with linear and logarithmic axes, surface and contour plots, Smith charts, and polar plots. The user may specify virtually every characteristic of the plots, including the text and position of labels. Full hardcopy capabilities are also supported. TONYPLOT includes animation features that permit viewing a sequence of plots in a manner showing solutions as a function of some parameter. The parameter can be varied under slider control, or frames can be looped continuously, a feature very helpful in developing physical insight. TONYPLOT supports production of hard copy plots on a wide range of printers.

3.3 Device physics

The simulation technique is a numerical one, which arrives at the terminal characteristics of the device by solving the partial differential equations describing the physics of the materials and the effects of potentials and heterojunctions on carrier transport. These equations include Poisson’s equation, the carrier continuity equations and the transport equations for each carrier (electrons and holes). These equations must be solved simultaneously and self consistently in each region of the device. Here, a brief description of the basic semiconductor equations involved is provided.
3.3 Device physics

3.3.1 Poisson’s equation

Poisson’s equation relates the electric field distribution in space of the charge and is given in one dimension by [173],

\[
\frac{dE(x)}{dx} = \frac{q}{\varepsilon_s}(p - n + N_d^+ - N_a^-) \tag{3.1}
\]

where \(E(x)\) is the electric field, \(\varepsilon_s\) the dielectric constant of the semiconductor, \(p\) and \(n\) are the carrier concentrations and \(N_d^+\), \(N_a^-\) are the concentrations due to ionised donor and acceptor impurities, respectively all of which can be a function of position. Clearly, the doping in each region of the device must be specified at the outset in order to describe the electric field everywhere within the device. An alternate form of the Poisson equation is obtained when the electric field \(E(x)\) in (3.1) is replaced by \(-dV/dx\) as given in (3.2). Generally both the equations are referred to as the Poisson’s equation.

\[
E(x) = -\frac{dV}{dx} \tag{3.2}
\]

3.3.2 Carrier continuity equations

The continuity equations for the electrons and holes are given by [172],

\[
\frac{\partial n}{\partial t} = \frac{1}{q} \nabla \cdot J_n + G_n - R_n \tag{3.3}
\]

\[
\frac{\partial p}{\partial t} = -\frac{1}{q} \nabla \cdot J_p + G_p - R_p \tag{3.4}
\]

where \(n\) and \(p\) are the electron and hole concentrations, \(J_n\) and \(J_p\) are the total electron and hole current densities, \(G_n\) and \(G_p\) are the generation rates for electrons and holes due to optical absorption, \(R_n\) and \(R_p\) are the net recombination rates for
3.4 Physical models

electrons and holes, respectively, and \( q \) is the magnitude of the electron charge. The rates of recombination, \( R_n \) and \( R_p \), are proportional to the excess carrier concentration and are given by \( R_n = (n-n_0/\tau_n) \) and \( R_p = (p-p_0/\tau_p) \) where \( \tau_n \) and \( \tau_p \) are lifetimes of the electrons and holes, respectively.

3.3.3 Transport equations

Equations 3.1, 3.3, and 3.4 provide the general framework for device simulation. However, further secondary equations are needed to specify particular physical models for: \( J_n \), \( J_p \), \( G_n \), \( G_p \), \( R_n \), \( R_p \). The current density equations, or charge transport models, are usually obtained by applying approximations and simplifications to the Boltzmann transport equation. These assumptions can result in a number of different transport models such as the Drift-Diffusion Model, the Energy Balance Model or the Hydrodynamic Model. The choice of the charge transport model will then have a major influence on the choice of generation and recombination models.

The simplest model of charge transport that is useful is the Drift-Diffusion Model. This model has the attractive feature that it does not introduce any independent variables in addition to electrostatic potential and electron concentration. Until recently, the Drift-Diffusion Model was adequate for nearly all devices that were technologically feasible. However, the drift-diffusion approximation becomes less accurate for smaller feature sizes. More advanced Energy Balance and Hydrodynamic Models have become the preferred choice for simulating deep submicron devices. ATLAS provides both the Drift-Diffusion and the advanced transport models, which will be presented in the next section, where the physical models used in the simulations are discussed. Other models incorporated in the simulations include ones for the carrier generation-recombination mechanism, velocity saturation effects at high electric fields, which will also be covered in the later sub-sections.

3.4 Physical models

A great deal of analysis involving the transferred-electron effect in InGaAs showing its superiority as a highly efficient Gunn diode has been performed by Kowalsky et
3.4 Physical models

al. [174-178]. The key reasons include the good lattice-matching with InP substrates, high low-field electron mobility (12×10³ cm²/Vs), high peak electron velocity (2.9×10⁷ cm/s) and low threshold electric field (4 kV/cm) which minimise power dissipation. The InAlAs/InGaAs HEMTs lattice-matched to the InP substrate have shown superior microwave and low-noise characteristics over GaAs-based pHEMTs due to the higher two-dimensional electron gas (2DEG) density and the large conduction band discontinuities (ΔEC) between the Schottky contact layer and the channel layer [179-181]. For these reasons, the SSD is based on an In0.53Ga0.47As/In0.53Al0.47As heterostructure, where a 2DEG is formed at the heterointerface, shown in Fig. 3.2. The physical models invoked in the simulations are chosen for carrier transport in the active In0.53Ga0.47As layer since it is the 2DEG layer that largely determines the electronic properties of the device [50, 155].

![Figure 3.2 Schematic of the SSD device based on an In0.53Ga0.47As/In0.53Al0.47As heterostructure, where a 2DEG is formed at the heterointerface [158].](image)

### 3.4.1 Carrier statistics and transport

Within a semiconductor, the thermal equilibrium of electrons at temperature TL obeys Fermi-Dirac statistics. The electron occupation probability for a given energy is given as [172]:

\[
f(E) = \frac{1}{1 + \exp\left(\frac{E - E_F}{kT_L}\right)} \tag{3.5}
\]

If \(E - E_F \gg kT_L\) (non-degeneracy) then the equation can be approximated as:
3.4 Physical models

\[ f(E) = \exp \left( \frac{E_f - E}{kT_L} \right) \]  \hspace{1cm} (3.6)

The above approximation is known as the Boltzmann statistic. The Boltzmann approximation simplifies the calculations whilst yielding satisfactory results. It was used in the model that minimized convergence issues as compared to the Fermi-Dirac statistics. In ATLAS the use of the Boltzmann statistic has been set at its default implementation [172].

The transport models, Drift-Diffusion and Energy Balance Models from ATLAS were harnessed for the modelling of the Gunn diode. The equations governing these models are of essential knowledge and are detailed below. In the Gunn diode model the carriers were selected as electrons and only electrons equations are presented.

The Drift-Diffusion Transport Model

Derivations based upon the Boltzmann transport theory have shown that the current densities in the continuity equations may be approximated by a Drift-Diffusion Model. In this case the current density for electrons is expressed in terms of the quasi-Fermi level \( \phi_n \) as:

\[ J_n = q\mu_n n \nabla \phi_n \]  \hspace{1cm} (3.7)

where \( J_n \) is the electron current density, \( q \) is the electron charge, \( n \) is the electron concentration and \( \mu_n \) is the electron mobility. The quasi-Fermi level is then linked to the carrier concentration \( (n) \) and the potential \( (V) \) through the Boltzmann approximation:

\[ n = n_{ie} \exp \left[ \frac{q(V - \phi_n)}{kT_L} \right] \]  \hspace{1cm} (3.8)

where \( n_{ie} \) is the effective intrinsic concentration and \( T_L \) is the lattice temperature. The above may then be re-written to define the quasi-Fermi potential:
3.4 Physical models

\[ \phi_n = V - \frac{kT_L}{q} \ln \frac{n}{n_{ie}} \]  \hspace{1cm} (3.9)

By substituting this equation into the current density expression, the following adapted current relationship is obtained;

\[ J_n = qD_n \nabla n - qn \mu_n \nabla V - \mu_n n \left( kT_L \nabla \left( \ln n_{ie} \right) \right) \]  \hspace{1cm} (3.10)

The final term accounts for the gradient in the effective intrinsic carrier concentration, which takes into account the bandgap narrowing effects. The effective electric field is normally defined as,

\[ E_n = -\nabla \left( V + \frac{kT_L}{q} \ln n_{ie} \right) \]  \hspace{1cm} (3.11)

which allows the more conventional formulation of drift-diffusion equation to be written as:

\[ J_n = qn \mu_n E_n + qD_n \nabla n \]  \hspace{1cm} (3.12)

It should be noted that this derivation of the Drift-Diffusion Model has tacitly assumed that the Einstein relationship holds. In the case of Boltzmann statistics this corresponds to the diffusion coefficient \( D_n \) expressed as:

\[ D_n = \frac{kT_L}{q} \mu_n \]  \hspace{1cm} (3.13)

If Fermi-Dirac statistics are assumed, it becomes:
3.4 Physical models

\[
D_n = \frac{\left(\frac{kT_L - \mu_n}{q}\right)F_{1/2} \left\{ \frac{1}{kT_L} (\varepsilon_{F_n} - \varepsilon_C) \right\}}{F_{-1/2} \left\{ \frac{1}{kT_L} (\varepsilon_{F_n} - \varepsilon_C) \right\}} \tag{3.14}
\]

where \( F_\alpha \) is the Fermi-Dirac integral of order \( \alpha \) and \( \phi_{F_n} \) is given by \(-q\phi_n\).

**Energy Balance Transport Model**

The fundamental limitation of the Drift–Diffusion model is its assumption that the electron temperature, \( T_n \), is equal to the lattice temperature, \( T_L \). With this constraint, parameters such as the impact ionisation rates, the carrier mobility and the drift velocity (\( \nu_d \)) are linked to a local electric field rather than the spatial variation of \( T_n \).

As a result, it is possible to severely underestimate the performance of a device at nanometre dimensions because velocity overshoot is neglected and impact ionisation overestimated [182, 183].

To make the drift–diffusion model more accurate, \( T_n \) must be made to deviate from \( T_L \). This can be accomplished by relating \( T_n \) to the average kinetic energy of an electron, which can be written as the sum of an electron's drift and thermal energies [184]:

\[
W_n = \frac{1}{2}nm_e^*\nu_d^2 + \frac{3}{2}nkT_n \tag{3.15}
\]

Then, electron temperature gradients are incorporated into the Drift–Diffusion model by creating an additional balance equation representing the rate of energy lost by an electron to the lattice [184]. This additional balance equation is shown below:

\[
\frac{\partial W_n}{\partial t} = -\nabla \cdot F_n + J_n \nabla V - \frac{3}{2}nk \left( \frac{T_n - T_L}{\tau_e} \right) + G_n - R_n \tag{3.16}
\]

\[
F_n = W_n \nu_d + nkT_n \nu_d - K_n \nabla T_n \tag{3.17}
\]
3.4 Physical models

where \(K_n\) stands for the thermal conductivity of an electron and \(F_n\) denotes the flux of energy between an electron and the lattice; \(\tau_e\) represents the energy relaxation time; \(i.e.,\) the time needed for the energy (temperature) distribution to reach steady state with the electric field [185]. To complete the modification of the drift–diffusion model, (3.12) is altered to include a dependence on \(T_n\) [184]:

\[
J_n = qn\mu_n E_n + qD_n\nabla n + n\mu_n k\nabla T_n
\]

(3.18)

The system of equations defined by (3.3) and (3.16) to (3.18) represents what is known as the Energy Balance Transport Model. The model can be activated in ATLAS by selecting \(HCTE.EL\) in the model statement. The Hydrodynamic Model encompasses the transport equations defined by the Energy Balance Model but differs in that the carrier mobilities are unaffected by carrier temperatures. The Hydrodynamic Model is applied by setting the proportionality constant from -1 (for Energy Balance) to 0 in the model statement.

3.4.2 Mobility

The carrier mobility is a parameter that characterises the drift motion of the carriers in the presence of an electric field. The electrons and holes are accelerated by electric fields, but lose momentum because of various scattering processes involving lattice vibrations, impurity ions, material interfaces and other imperfections [186, 187]. The effects of these microscopic phenomena are averaged and lumped together into the macroscopic parameter called the mobility, which is a function of the local electric field, doping concentration, and temperature.

In the Gunn diode, as in other electron carrier devices, the electrons are accelerated by the electric field. The carriers are almost in equilibrium with the lattice at low field values thus the electron mobility has a characteristic low-field value subject to phonon and impurity scattering. However, the carriers are no longer in equilibrium at high electric fields and are consequently subjected to a wider range of scattering processes. Therefore, the drift velocity no longer increases linearly and becomes saturated or decreases, which is termed as saturation velocity. These effects
3.4 Physical models

have to be accounted for using the appropriate models [172]. ATLAS and its sub-engine Blaze provide various low-field and high-field models that can be used depending on the device and its operation. The models used for the simulations will now be discussed.

**Low-field mobility**

Five distinctive models and conditions are available in ATLAS to define the low field electrical mobility in InGaAs [172]:

- The low-field mobility parameter, $MUN$ (cm$^2$/V.s), can be defined from the ATLAS lookup table for different materials. The value is defined in the mobility statement.
- The concentration based mobility model $CONMOB$ can be used. The values for low field mobility at 300K are found from the ATLAS look-up table for the doping concentration in each region.
- In order to relate the low-field electron mobility with impurity concentration and temperature, the $ANALYTIC$ and $ARORA$ models, may be used.
- A carrier-carrier scattering model, $CCSMOB$, which relates the low-field mobility to both carrier concentration and temperature, can be used.
- The unified low-field mobility model, $KLAASSEN$, relates the low field mobility to lattice, carrier-carrier, donor scattering, and temperature.

In formulating an unknown structure which can consist of multilayer semiconductor materials, whose electrical properties are unknown, it will be useful to carefully consider each low mobility model that may apply to each layer or material; so that the mobility of the overall device may be predicted accurately. In the case of the Gunn device used in this work, the 2DEG properties of the heterostructure based on InGaAs/InAlAs are widely known [188-190]. The simplest model of defining the low-field mobility parameter using $MUN$ is applied with a value of $12\times10^3$ cm$^2$/V.s [189, 191]. The high-field model is discussed below.
3.4 Physical models

Parallel Electric Field-Dependent Mobility

Two types of electric field-dependent mobility (FLDMOB) models are available in ATLAS to model electron transport at high-fields. The models are named as the Standard Mobility model and Negative Differential Mobility model, both of which contain appropriate default parameter values for different materials. The models are defined by specifying EVSATMOD along with the FLDMOB in the model statement. The EVSATMOD specifies which parallel field dependent-mobility model should be used for electrons, which are defined as [172]:

- EVSATMOD = 0 allows the application of standard mobility model
- EVSATMOD = 1 implements the negative differential mobility saturation model.

The Barnes Negative Differential Mobility model [192] was used for high electric field transport. The electron velocity, as a function of electric field, is given by:

\[ \mu(E) = \frac{\mu_0 + \frac{v_{sat}}{E} \left( \frac{E}{E_{CRIT}} \right)^\Gamma}{1 + \left( \frac{E}{E_{CRIT}} \right)^\Gamma} \]  

(3.20)

where \( v_{sat} \) is the electron saturation velocity, \( \mu_0 \) is the low field electron mobility, \( E_{CRIT} \) is the critical electric field in V/cm, \( \Gamma \) is a fitting constant equal to 2 and \( E \) is the electric field. The critical electric field for InGaAs was defined as \( E_{CRIT} = 4.0 \times 10^3 \) [193]. For the device modelled at 300 K, a saturation velocity value of \( 2.0 \times 10^7 \) cm/s was used [190, 193, 194].

3.4.3 Carrier generation-recombination models

The processes responsible for carrier generation and recombination fall under following six main categories:
3.4 Physical models

- Phonon transitions
- Photon transitions
- Auger transitions
- Surface recombination
- Impact ionisation
- Tunnelling

The six generation-recombination mechanisms are represented using different models in ATLAS. For Gunn diode modelling, the Shockley-Read-Hall (SRH) recombination (phonon transition) has been used.

**Shockley-Read-Hall (SRH) recombination**

Phonon transition within a semiconductor is due to a trap or defect within its forbidden gap. The theory of phonon transition was discovered first by Shockley and Read [195], and then later by Hall [196]. The SRH recombination model is implemented as follows:

\[
R_{SRH} = \frac{pn - n_{ie}^2}{\tau_p \left[ n + n_{ie} \exp \left( \frac{E_{TRAP}}{kT_L} \right) \right] + \tau_{N0} \left[ p + n_{ie} \exp \left( \frac{-E_{TRAP}}{kT_L} \right) \right]} \tag{3.21}
\]

where \( \tau_{N0} \) and \( \tau_p \) are the electron and hole life times, the value of which was taken to be \( 1.0 \times 10^{-8} \) s for \( In_{0.53}Ga_{0.47}As \) [197, 198]. \( E_{TRAP} \) indicates the trap energy and intrinsic Fermi level difference in electron volts. Its default value (\( E_{TRAP} = 0 \)) was used to correspond to efficient recombination of one trap layer present in the material. \( T_L \) is the lattice temperature in degrees Kelvin. The model was activated by defining \( SRH \) in the model statement.

The full set of physical models used in the Gunn diode simulations are summarised in Table 3.1.
3.5 Model validation

To obtain the most accurate results, the simulated results using the models in ATLAS are calibrated to Monte Carlo simulations. Figure 3.3 shows one of the $I$-$V$ characteristics of SSDs using ATLAS simulation compared to its Monte Carlo simulation. The close match of the results validates the models and parametric values used.

**Table 3.1** Summary of the physical models used in the Gunn device simulations.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Model</th>
<th>Syntax</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transport</td>
<td>Energy Balance</td>
<td>$HCTE.EL$</td>
</tr>
<tr>
<td>Mobility</td>
<td>Parallel Electric Field-Dependent</td>
<td>$FLDMOB$</td>
</tr>
<tr>
<td>Carrier generation-recombination</td>
<td>Shockley-Read-Hall</td>
<td>$SRH$</td>
</tr>
</tbody>
</table>

![Figure 3.3 I-V characteristics compared between ATLAS simulation and the Monte Carlo simulation [144, 145] of two SSDs with channel width $W = 50$nm having different channel lengths: (i) $L = 0.1$μm and (ii) $L = 1.0$ μm.](image-url)
3.5 Model validation

Further to this, a time-transient ATLAS simulation was also compared to the previously published Monte Carlo result. The oscillation obtained in ATLAS, Fig. 3.4a, is matched to Fig. 3.4b [50]. Parameters and conditions applied to both simulations were kept constant.

![Figure 3.4](image)

**Figure 3.4** Current response of a SSD with channel length $L=1.25$ μm and channel width $W=60$ nm using (a) ATLAS and (b) Monte Carlo simulations.

The excellent agreement of the two results fully validates the physical reliability of the configured modelling technique. Based on this, Gunn diode simulations are conducted on SSD structures to verify the Gunn Effect taking place in the devices. This is done by investigating the electron dynamics within the conduction channels as they are conveniently plotted using TONYPLOT in ATLAS. Also, the geometrical designs of the SSDs can be varied and studied with less difficulty than using Monte Carlo simulations. In the next chapter, the impact of geometry and other critical parameters on Gunn oscillations in the SSD are reported.
4

Planar Gunn diodes

4.1 Introduction

In Chapter 2, the planar SSD was introduced citing its numerous functions such as rectification and THz detection. The capability to emit THz radiation was also highlighted since it has been investigated using Monte Carlo simulations. In this chapter, the radiation characteristics are studied in detail using physical modelling.

The simulation package ATLAS is used and the models developed were reported in Chapter 3, including a validation with the Monte Carlo results. The simulated SSD device is based on the $\text{In}_{0.53}\text{Ga}_{0.47}\text{As} / \text{In}_{0.53}\text{Al}_{0.47}\text{As}$ heterostructure. The conduction channel is designed to be $L = 1.2 \, \mu\text{m}$ in length and $W = 80 \, \text{nm}$ in width and the other geometric parameters are defined in Fig. 4.1. To account for the fixed positive charge of the doping layer, a virtual net background doping $N_v = 1.0 \times 10^{17} \, \text{cm}^{-3}$ (without impurity scattering) is assigned to the $\text{In}_{0.53}\text{Ga}_{0.47}\text{As}$ layer. The simulations are carried out at room temperature condition.

4.2 DC Conditions

Influence of interface-charge density on $I-V$ characteristics

It is well-documented that the interface-charge density on the semiconductor-insulator boundaries plays a principal role in controlling the conduction current in the
4.2 DC Conditions

Figure 4.1 Geometry of the simulated SSD in microns, not drawn to scale.

The nonlinearity of the current-voltage relationship that is characteristic of the SSD is verified first using different values for the charge density. Later, a simulated I-V characteristic is compared to one that is experimentally obtained to identify a reasonable value for subsequent studies. Figures 4.2a, b and c show the I-V characteristics of the diode with different interface-charge densities $\sigma = 0$, $4 \times 10^{11}$ and $6 \times 10^{11}$ cm$^{-2}$. The results show the characteristic asymmetric behaviour of a conventional diode substantiating the SSD working principle and experimental results [46]. In the absence of surface charges (Fig. 4.2a), the asymmetric behaviour is less pronounced. When $\sigma = 4 \times 10^{11}$ cm$^{-2}$ (Fig. 4.2b), the channel is pinched off for the reverse bias, producing a zero threshold device. For a visual representation of the depletion profiles, Fig. 4.5a can be referred to. Finally, the introduction of more charges at the interfaces (Fig. 4.2c) causes electron-electron repulsion in the channel, depletes the channel of electrons and requires a higher threshold voltage $V_{th} = 3$ V to switch the diode on (Fig. 4.2c).

Experimentally, the interface-charge present at the semiconductor-dielectric boundaries can vary considerably depending on the fabrication techniques such as etching methods. The I-V characteristic of an $In_{0.53}Ga_{0.47}As$ based SSD fabricated using electron beam lithography and wet chemical etching is shown in Fig. 4.3. The dimensions of the SSD are $L = 1.5$ μm and $W = 140$ nm. The simulation of the same structure using ATLAS produced a close match using $\sigma = 6.5 \times 10^{11}$ cm$^{-2}$. Thus, $\sigma = 2 \times 10^{11}$ to $10 \times 10^{11}$ cm$^{-2}$ is an appropriate range to use in the simulations and will be employed for characterisation of the devices.
4.3 Electron distribution in the device

This section shows the advantage of being able to study devices using physical modelling since the processes occurring within the device cannot be appreciated experimentally. The electron distribution in the channel is compared for the different conditions of interface-charge density used in the previous section to gain an understanding of the underlying reasons for the results obtained. The electron density at zero bias in the device along the channel is plotted against the distance along the line in Fig. 4.4. From the electron density shown in Fig. 4.4a, the absence of interface-charge density leaves the channel open for electron transport both for positive and negative biases, thus the conduction $(I-V)$ is symmetrical. In Fig. 4.4b, the introduced interface-charge density repels the electrons away from the channel thus causing a drop in the density of the conducting electrons, more so in the region of greater trench surface area at the entrance of the channel nearer to the cathode; this enables the channel to become conductive in the forward bias and resistive in the negative bias. However for a greater interface-charge density in 4.4c, the repulsion is so great the channel becomes drastically devoid of electrons (about 8 orders in magnitude), Fig. 4.4d. A very high electric field would be necessary to propagate the electrons through the channel, hence the high threshold voltage as shown in Fig. 4.2.

Figure 4.2 $I-V$ characteristics of an SSD with $L = 1.2 \, \mu m$, $W = 80 \, nm$ for different interface-charge densities.
4.3 Electron distribution in the device

\[ \text{L = 1.5 } \mu \text{m W = 450 nm} \]

![Comparison of the experimental I-V characteristic of a SSD with L = 1.5 \( \mu \)m and W = 450 nm with the simulated I-V performed using ATLAS.](image)

**Figure 4.3** Comparison of the experimental I-V characteristic of a SSD with \( L = 1.5 \) \( \mu \)m and \( W = 450 \) nm with the simulated I-V performed using ATLAS.

![The electron concentration per cm\(^3\) at 0V in the channel; altered by the introduction of interface-charge density.](image)

**Figure 4.4** The electron concentration per cm\(^3\) at 0V in the channel; altered by the introduction of interface-charge density.
4.3 Electron distribution in the device

Influence of applied bias on electron density

By fixing \( \sigma = 4 \times 10^{11} \text{ cm}^{-2} \), the variation in electron density is observed for positive and negative bias to understand the nonlinear behaviour of the device. Though the mechanism underlying the phenomenon is well-established, physical modelling has yet to be performed showing top-view of the electron density contours in the device. Fig. 4.5a shows the effect of applying a positive and negative bias on the electron density in the channel. When the channel is negatively biased (-0.1V), the electron density reduces further, verifying the induced negative charges in the channel which form a barrier to electron flow, effectively pinching off the channel. Conversely, a positive bias (0.1V) shows an increase in the electron density proving that more electrons are able to traverse the channel, overcoming the depletion with sufficient carriers, therefore opening up the channel. The pinched-off and open channel regions for biases up \( \pm 1 \text{ V} \) are also shown with their corresponding profile lines featured in Fig. 4.5b.

Influence of electron density on the \( I-V \) characteristic

The current passing through the SSD is determined by the electron density and velocity in the nanochannel. Monte Carlo simulations performed on the active InGaAs 2-DEG layer show that both the average electron density and velocity in the channel vary significantly as a function of applied bias but the electron density changes much more strongly with the bias voltage than the velocity [154], Fig. 4.6b.

Although a pronounced negative differential electron velocity is observed for voltages above 0.8 V, the \( I-V \) characteristic of the SSD in Fig. 4.6a still shows a continued increase in current because the accumulation of electrons dominate. The bias-dependent variation of electron density is therefore reported to be the main contributor to the diode-like characteristic of the SSD and this finding is verified by the ATLAS model for the InGaAs device. The average electron density in the channel is plotted in Fig. 4.7. The similar result confirms that the dominant factor determining the \( I-V \) characteristic of the SSD is the electron density in the channel.
4.3 Electron distribution in the device

Figure 4.5 (a) Contour plots and (b) profile lines of electron density for incremental positive and negative biases.
4.3 Electron distribution in the device

Figure 4.6 (a) Calculated I-V characteristic of an InGaAs SSD with a $L = 1.0 \, \mu m$ and $W = 70 \, nm$ and (b) average electron density and velocity in the InGaAs SSD channel versus applied bias [154].

Figure 4.7 Average electron density versus applied bias simulated by ATLAS for the SSD in Fig. 4.1 ($L = 1.2 \, \mu m$ and $W = 80 \, nm$).
4.4 Gunn oscillations

The principle of the Gunn effect, presented in Chapter 2, is attributed to an intervalley transfer of electrons at a critical electric field. The difference in effective masses and mobilities of the electrons in these valleys can cause significant slowdown of their velocity, culminating in the formation of domains. This phenomena is illustrated here, using electron dynamics in the device to show that the oscillations produced are indeed due to the Gunn effect.

Figures 4.8a and b show the current responses obtained when the voltage applied is incremented in small steps. Sustained oscillations arise from 1.5 and 2.0 V for channel widths of 80 and 100 nm respectively. Their respective frequencies are 0.181 and 0.120 THz. It is also apparent that their amplitudes differ significantly. This initial discovery of a contrast in oscillating characteristics in different geometries led to the comprehensive study of device structures and their corresponding oscillations. This was followed by a detailed analysis of electron transport in the devices to determine the reasons for their behaviour.

4.4.1 Domain profiles

The current oscillation from Fig. 4.8b at 2.5 V is analysed by segmenting one oscillation period into 6 time moments (Fig. 4.9a); for the purpose of studying the dynamic behaviour of the electrons. Figures 4.9b and c show the electron density and electric field distributions along the centre of the channel during these 6 time moments. It is easily observed in Fig. 4.9b that a charge domain consisting of an electron accumulation (negative charge) and an electron depletion region (positive charge) form adjacent to each other at the beginning of the channel; which then grows and propagates along the channel. When the domain terminates at the end of the channel, another domain forms at the beginning of the channel and the process repeats resulting in a periodical current. It is also noteworthy that the domain initiation takes place just outside the location of the vertical trench. This is because of the electron depletion within the vertical trench area owing to greater surface depletion and hence the presence of interface-charge density states. In Fig. 4.9c, the corresponding electric field domains as a result of the charge domains are shown. The relationship between the two domains validates the classical Gunn phenomenon.
4.4 Gunn oscillations

Figure 4.8 Current response from the SSD when biased in incremental bias steps for (a) $L = 1.2 \ \mu m$, $W = 80 \ \text{nm}$; and (b) $L = 1.2 \ \mu m$, $W = 100 \ \text{nm}$.

In order to attribute specific electron distribution in the channel to its corresponding current value, the six moments are reviewed individually with reference to Fig. 4.9c. Moment 1 where the current first drops from its peak value is the moment when a new domain forms (an electric field peak) at the entrance of the channel. As the electric field begins to build at this end, the supplied voltage is dropped, increasing across this section and the current continues to fall (moments 2-3). At moment 4, the electric field domain enters the second half of the channel and is now closer to the anode allowing the current to pick up. When the domain reaches the end of the
channel, all of the voltage supplied is dropped there, resulting in the peak current being registered (moment 6). Fig. 4.9d shows the top-view electric field contour plots illustrating the domains formation, propagation and termination.

\[
\begin{align*}
L &= 1.2 \text{ µm} \quad W = 100 \text{ nm}
\end{align*}
\]

**Figure 4.9** (a) Current oscillation from Fig. 4.8b at \( V = 2.5 \text{ V} \) showing 6 time moments in one period, (b) electron density line profiles along the centre of the channel, (c) electric field line profiles along the centre of the channel; and (d) electric field contour plots showing the domain formation, growth, propagation and termination in the channel. The vertical lines in (b) and (c) indicate the beginning and end of the channel.
4.4 Gunn oscillations

4.4.2 Threshold bias

As mentioned in the previous section, the interface-charge density can influence the spatial electron density in the device. Since the depletion of electrons is greater around the location of the vertical trenches, \(0.3 < X < 0.5\), the electric field is higher locally as compared to either sides of the vertical trench, Fig. 4.10b. Therefore, within this local region, the electrons attain the required energy to transfer to the satellite valley, slow down owing to their new (heavier) effective electron mass and form charge domains.

The average electric field across this region is \(~ 4\ \text{kV/cm}\) for 1.8 V; corresponding to the Gunn threshold electric field for InGaAs and matches the oscillation threshold found in Fig. 4.8b. Further evidence of the electron dynamics varying locally at the position of the vertical trenches is plotted in Figs. 4.10c and d, showing the energy...
and the velocity distribution respectively. Thus the peak electric field located at the entrance of the channel is a favourable condition for domain formation. The effect is similar to notch doping where a lightly doped region is incorporated adjacent to the cathode of a Gunn diode to reduce 'dead space' hence improving the efficiency of operation [199]. Here, the SSD achieves intended similar effect of initiating domain formation near the cathode, simply by its asymmetrical geometry in a planar rather than vertical device structure.

4.4.3 Cessation bias

Besides originating at a threshold bias, oscillations also cease at a bias when the electrons reach their saturation velocity. This aspect is critical in order to realise the role of the device geometry in determining the bias range of Gunn oscillations. The biases at which oscillations cease for different channel dimensions can be appreciated from Fig. 4.11.

![Graph showing current oscillations with different biases](image)

Figure 4.11 The current oscillations showing the threshold and cessation biases for (a) the reference geometry: $L = 1.2 \, \mu m$, $W = 100 \, nm$, (b) the longer channel: $L = 1.8 \, \mu m$, $W = 100 \, nm$; and (c) a wider channel $L = 1.2 \, \mu m$, $W = 180 \, nm$. 
4.4 Gunn oscillations

The data of threshold and cessation biases for achieving Gunn oscillations in a range of channel lengths and channel widths are given in Tables 4.1 and 4.2 respectively. The threshold voltage was largely determined by the channel’s width whereas its length had a much lesser influence. But in the case of the oscillation cessation bias, lengthening the channel allows the oscillations to be sustained for a wider range of biasing, presenting a broader working range. Such details can be invaluable to experimentalists. Further information relating to the oscillation performances based on the applied biases will be discussed in the next section.

W = 100 nm

<table>
<thead>
<tr>
<th>Channel length (μm)</th>
<th>Threshold bias (V)</th>
<th>Cessation bias (V)</th>
<th>Oscillation bias range (V)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>2.0</td>
<td>3.0</td>
<td>1.0</td>
</tr>
<tr>
<td>1.2</td>
<td>2.0</td>
<td>3.5</td>
<td>1.5</td>
</tr>
<tr>
<td>1.8</td>
<td>2.5</td>
<td>6.0</td>
<td>3.5</td>
</tr>
<tr>
<td>2.5</td>
<td>2.5</td>
<td>7.5</td>
<td>5.0</td>
</tr>
</tbody>
</table>

Table 4.1 Working bias range for different channel lengths.

L = 1.2 μm

<table>
<thead>
<tr>
<th>Channel width (nm)</th>
<th>Threshold bias (V)</th>
<th>Cessation bias (V)</th>
<th>Oscillation bias range (V)</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>1.8</td>
<td>3.2</td>
<td>1.4</td>
</tr>
<tr>
<td>140</td>
<td>3.2</td>
<td>5.7</td>
<td>2.5</td>
</tr>
<tr>
<td>180</td>
<td>4.7</td>
<td>7.2</td>
<td>2.5</td>
</tr>
<tr>
<td>220</td>
<td>5.2</td>
<td>8.2</td>
<td>3.0</td>
</tr>
</tbody>
</table>

Table 4.2 Working bias range for different channel widths.

4.4.4 Factors affecting oscillating performance

To achieve high performance operation at THz frequencies, full optimisation of the device design is required. For this, the SSD device emitter performance is studied in terms of its frequency and amplitude derived from the Fourier transforms of the
current oscillations. The performance has been analysed with respect to three key parameters, namely channel length, channel width and interface-charge density since all these parameters can be adjusted by lithography and hence are relevant to practical applications.

### Channel-length dependence

The frequency and amplitude of the oscillations determined using Fourier transforms of the oscillations are recorded for different channel lengths in Fig. 4.12a. Since the threshold voltage for the longest channel is 2.5 V, the results for all the channel lengths are extracted at 2.5 V. The fundamental frequency is inversely proportional to the length of the channel, Fig. 4.12b. The current oscillation amplitudes however are fairly close to each other which suggest that they are relatively independent of the channel length.

![Figure 4.12](image)

**Figure 4.12** (a) Frequencies and amplitudes of the fundamental frequencies for increasing channel lengths, $W = 100$ nm, $\sigma = 4.0 \times 10^{11}$ cm$^{-2}$ and $V = 2.5$ V. The channel length has a stronger influence on the frequency than the amplitude of current oscillations. (b) The inverse relation of the fundamental frequency with the length of the channel.

Interestingly, it is from their current oscillations in Fig. 4.13 where contrasting features are observed for the different channel lengths. As the channel is made longer, the waveforms become less sinusoidal (insets). There appears an additional small peak in the oscillating current which becomes increasingly prominent. This directly influences the emission spectrum. A number of fairly strong harmonics up to 0.4 THz emerge as the channel length increases. In the channels exceeding 1.8 $\mu$m, at least four harmonic peaks, whose amplitudes are higher than 10% of the fundamental mode are present, whereas only two were present before in the same frequency range.
for the shortest channel \((L = 1.0 \, \mu m)\). Particularly, the second harmonic peak is greater in magnitude than the fundamental mode.

\[
W = 100 \, \text{nm}, \quad \sigma = 4.0 \times 10^{11} \, \text{cm}^2, \quad V = 2.5 \, \text{V}
\]

\[L = 1.0 \, \mu m \quad L = 1.2 \, \mu m \quad L = 1.5 \, \mu m \quad L = 1.8 \, \mu m \quad L = 2.0 \, \mu m \quad L = 2.5 \, \mu m\]

**Figure 4.13** Frequency spectra of the SSDs with increasing channel lengths from 1.0 to 2.5 \(\mu m\); and their waveforms are shown as insets. A greater number of harmonics within a narrow frequency range become possible owing to the nature of the waveforms.

In order to determine the reason behind the findings, the electron density domain dynamics that takes place within the longer channel \((L = 1.8 \, \mu m)\) is examined in Fig. 4.14. This is compared to the domain dynamics already discussed for the shorter channel \((L = 1.2 \, \mu m)\) in Fig. 4.9. The period of oscillation is separated into three
4.4 Gunn oscillations

sections. The first is the section where the current falls. As identified before, this is when the domains initiate at the beginning of the channel and grow in size as they propagate the channel. In section two, the current dips again. This is when a sizeable domain has been achieved in the middle of the channel, Fig. 4.14b. It would require a substantial bias for the domains to acquire the required energy for efficient propagation through this middle section of the channel. However, the supplied bias is only 2.5 V and furthermore, the channel length is longer. For these reasons, the domains cannot reach the end quickly enough, resulting in the drop in current at the anode. When the domains get to the end, the current quickly establishes its peak value. For this reason, the additional small peak/dip is observed, giving the oscillation a smaller frequency output and also a strong double harmonic spectrum. This explanation is verified by increasing the voltage supply to the device. The result plotted in Fig. 4.11b shows clearly that a higher bias in the longer channel (L = 1.8 μm) would ensure more efficient domain propagation. In this instance, as the bias is increased, the waveforms become more sinusoidal and the resulting frequency is also higher, i.e. 0.125 THz at 4.5 V compared to 0.087 THz at 2.5 V.

Channel-width dependence

The channel width dependence is shown in Fig. 4.15. Contrary to the length dependence, the width of the channel shows itself to be a significant determinant of oscillation amplitude but with a minimal effect on the frequency. In Fig. 4.16, the current waveforms from devices of different channel widths and their corresponding Fourier transforms are recorded. The possibility of engineering the device dimensions to create a greater number of significant higher harmonics in the required frequency bands becomes feasible. For example, in the device with W = 200 nm, as many as eight harmonics with considerable amplitudes are created reaching 1 THz.
4.4 Gunn oscillations

$L = 1.8 \, \mu m \quad W = 100 \, nm$

![Figure 4.14](image)

**Figure 4.14** The current oscillation for a long channel ($L = 1.8 \, \mu m$) analysed in three sections by plotting their electron density distributions. The dip in section (ii) occurs owing to the increased domain propagation distance and the insufficient bias (2.5 V) to drive the domains efficiently through the middle section of the channel.

![Figure 4.15](image)

**Figure 4.15** Frequencies and amplitudes of the fundamental frequencies for increasing channel widths, $L = 1.2 \, \mu m$ and $\sigma = 4.0 \times 10^{11} \, cm^{-2}$. The results are extracted from their respective threshold voltages which increase proportionately from 1.5 to 6.5 V. The channel width has a stronger influence on the amplitudes of oscillations than their frequencies.
4.4 Gunn oscillations

To realise how the width of the channel can alter the amplitudes of oscillations, the domains formed in the channels are revisited. From Fig. 4.17, the domains propagating in the wider channel ($W = 180$ nm) are greater in magnitudes as compared to the narrower one ($W = 100$ nm). This is a reasonable occurrence since domain sizes are determined by electron densities[20]. Hence, wide channels which contain a higher number of electrons per unit length will generate larger domains throughout the channel, giving rise to the high amplitudes in current oscillations. Since the current oscillation amplitude is highly tuneable by this method, this could facilitate power tuning when delivered to a specific load.

Figure 4.16 Frequency spectra of the SSDs with increasing channel widths from 100 to 200 nm; and their waveforms are shown as insets. The magnitude of the fundamental harmonic peak represents the amplitude of oscillations. A greater number of harmonics within a broad frequency range is achievable in wider channels.
4.4 Gunn oscillations

Although the individual parameters of length and width of the SSD channel assist in enhancing frequency and current oscillation amplitude respectively, the knowledge gained in this section can be utilised to develop structures and devices to suit a large variety of applications; for example, when a broad spectrum of THz radiation is needed or when a large peak power radiation is desired in a particular frequency band.

![Graphs showing electron and electric field density](image)

**Figure 4.17** (a) A greater number of electrons are contained in the domains of channels that are wider (W = 180 nm). (b) The electric field domains are correspondingly larger meaning a larger bias is required for them to traverse the channel.

### Interface-charge density dependence

Apart from geometries and their implications, materials employed for Gunn oscillations can have very different interface-charge densities at the etched interfaces depending on fabrication techniques, particularly different wet and dry etching methods and surface passivation techniques. This can influence the DC characteristics due to its direct effect on the depletion depth around the etched trenches. Its impact on the THz emission will add valuable information to previously acquired knowledge in the previous sections and will be considered here. The value of the interface-charge density at the semiconductor interfaces is increased progressively from $2 \times 10^{11}$ to $10 \times 10^{11}$ cm$^{-2}$ (Fig. 4.18). As the interface-charge number densities are increased, the channels are made wider to maintain approximately the same current and derive oscillations at the same threshold voltages.
4.4 *Gunn oscillations*

Remarkably, the highest interface-charge density enhanced the fundamental oscillating frequency from 0.12 to 0.22 THz, which is approximately 80% higher. However, a trade-off in the amplitude occurs as expected from the studies concluded previously. In Fig. 4.19c, the emission spectrum shows the second and third harmonics reaching 0.45 and 0.66 THz respectively. These are exceptionally high frequencies achieved from a device of high interface-charge density. Furthermore, a peak reaching 1 THz is achieved by only the fourth harmonic mode whereas higher modes of sixth and eighth were required in channel length and width modulations respectively.

![Figure 4.18 Frequencies and amplitudes of the fundamental frequencies for increasing interface-charge densities, L = 1.2 μm and V = 2.5 V. The widths of the channels are increased to accommodate similar average currents in the channels. A strong influence on the frequency of emission is observed.](image)

To understand why the frequency enhancement takes place, the electron domain dynamics is plotted for three different interface-charge density values in Fig. 4.20. For $\sigma = 2 \times 10^{11} \text{ cm}^{-2}$ in Fig. 4.20a, the domain formation is still close to the end of the vertical trench. However, as the value for the interface-charge density increases, the X position where the domain initiates, shifts towards the right, further into the channel (Figs. 4.20b and c). This can be attributed to the increased depletion area at the entrance of the channel where the vertical trenches are located as a consequence of the increased number of surface charges. The electrons then accumulate only
4.4 Gunn oscillations

beyond the severely depleted region thus forming a domain further inside the channel. As a result, the time transit of the domain is shorter due to the shortening of effective channel length and the emission frequency increases at high interface-charge densities. The presence of surface charges becomes an advantage rather than a disadvantage in a device like the SSD, since in many conventional devices, these charges are attempted to be eliminated by using encapsulation materials and/or passivation processes. As such, fabrication of the device becomes a lot less critical since it is easier to induce surface charges than to eliminate them.

Figure 4.19 Frequency spectra of the SSDs with increasing interface-charge densities from $2\times10^{11}$ to $10\times10^{11}$ /cm$^2$ and their waveforms are shown as insets. The high interface-charge density device produces the highest fundamental frequency and the fourth harmonic reaches 0.9 THz.
4.4 Gunn oscillations

Figure 4.20 The electron density domain initiation profiles show that the domain forms further inside the channel as the interface-charge number density is increased. The effective propagation distance is reduced and the frequency is enhanced.

Bias dependence

In Tables 4.1 and 4.2, the oscillation threshold and cessation biases were reported as function of channel length and width. In this section, the oscillation performance at the two ends of the bias range is compared and discussed. Fig. 4.21a shows the oscillation waves for these biases for three devices. Keeping the reference device dimension to be L = 1.2 μm and W = 100 nm, the effect on the Fourier emission spectra is studied in Figs. 4.21b, c and d. This is carried out at both the threshold bias and just before the cessation bias. This will give an indication of which oscillation should be extracted from the device, be it at the lower or the higher bias. For the reference device in Fig. 4.21b, the working bias range is just 1 V and the results are not much different, except that higher amplitudes of oscillations are achievable. For the longer channel in Fig. 4.21c, the emission becomes more monochromatic at the higher bias of 7 V compared to 2.5 V, the reasons for which have already been established in the section on length dependence. The amplitude of oscillation is also lower since the higher energy supplied to the electrons drive the domains to the end quicker before being able to build to a sizeable domain. In the third device with the wider channel in Fig. 4.21d, the amplitudes and fundamental frequencies experience minimal change but the major effect is on the number of harmonics possible within
4.4 Gunn oscillations

the frequency range. Reiteratively, another parameter like the bias is able to have an effect on a multitude of performance enhancing characteristics for Gunn oscillations.

Figure 4.21 (a) The waveforms of the devices showing their relative frequencies and amplitudes of oscillations with respect to geometry and their Fourier transformed emission peaks at (i) the threshold bias and (ii) the highest bias before cessation for (b) the reference geometry: \( L = 1.2 \, \mu m, \, W = 100 \, nm \), (c) the longer channel: \( L = 1.8 \, \mu m, \, W = 100 \, nm \); and (d) a wider channel \( L = 1.2 \, \mu m, \, W = 180 \, nm \).
4.4 Gunn oscillations

Harmonics of the Gunn oscillation

The data of higher harmonics generated by the SSDs as a function of geometrical parameters and interface-charge density is summarised in Fig. 4.22. This information can act as an additional tool for experimentalists to analyse and design devices for applications. For example, in Fig. 4.22b, the oscillations in narrow channels below 100 nm are strictly monochromatic which explain the absence of higher harmonics in these structures.

Figure 4.22 First, second and third harmonics data for the geometrical variations in (a) channel length, (b) channel width; and (c) interface-charge density.
4.5 Discussion and conclusions

By means of simulations, it has been shown that the SSD structure is able to function as a high-frequency Gunn diode, where a variety of geometries are possible. The active semiconductor channel dimensions, like length and width provide a good degree of frequency and magnitude selection and tuneability. Fundamental oscillation frequencies of 0.2 THz are achievable in the shortest channel simulated. The emission frequency can be increased by narrowing the channel. However, wider channel widths are better for improving the emission peak magnitude, directly related to derivable emission power. Furthermore, the interface-charge density in a device is found to influence the oscillations significantly. For a charge density of $1 \times 10^{12} \text{ cm}^{-2}$, the radiation frequency of higher harmonics could reach 0.7 THz. Thus, it is expected that a combination of carefully selected parameters, and/or using materials like GaN which emit in the higher frequency range, can enable generation of a broad spectrum with higher harmonics beyond 1 THz, which has so far been difficult to achieve in solid-state devices. The highlight of the findings is the determination of working bias ranges for different geometries; since it would be time-consuming to study this by employing trial and error methods in the lab.

The ease of fabrication and a flexibility in operation promises to unlock some of the difficulties posed by present THz technology. The findings here may guide experimentalists to design and realise such novel Gunn diodes, particularly on achieving very high fundamental frequencies and pronounced higher harmonics.
5

Planar nano-memory devices

5.1 Introduction

The colossal growth in the number of portable electronic devices today is attributed to a variety of emerging electronic technologies, with memory devices being at the forefront. They provide solutions for high-performance and high-density information storage and retrieval. Particularly, in nonvolatile memory devices, such as floating gate field-effect transistors, data retention exceeds ten years or more. However, boundaries never seem to cease as higher speeds and low power consumption demands are constantly placed on research. The technology most promising to meet these requirements relies heavily on semiconductor solid-state devices since they demonstrate higher speeds at a much lower cost, hence receiving a lot of attention [200-205].

Owing to the advantages of metal-oxide-semiconductor (MOS) technology in device manufacturability and miniaturisation, most ultra-large-scale integrated (ULSI $\geq 10^7$ transistors on a chip) memory circuits are currently made using MOS memories [17]. Generally, the memories are classified under two categories in terms of their volatility and access speed: as volatile random-access memories (RAMs) and nonvolatile non-RAMs. A nonvolatile memory (NVM) does not require any power for the maintenance of the data (retention) whereas a volatile memory needs to be powered up in order to retain information. One of the most widely used type
of NVM devices is the flash memory device which uses a floating gate as the charge-storage layer. Its simple architecture and massive memory capacity make it the popular choice for many applications. Being single transistor-based, it is more scalable than other types of memory devices. The main memory in personal computers, video game consoles and smartphones is the Dynamic Random-Access Memory (DRAM). DRAM stores each bit of data in a separate capacitor within an integrated circuit. Since capacitors leak charge, the information eventually fades unless the capacitor charge is refreshed periodically. Similar to flash memories is its structural simplicity: only one transistor and a capacitor are required per bit. This allows DRAM to reach very high densities. Details on the working mechanisms for flash and DRAM technologies will be discussed in 5.2.

Though simple and single transistor-based, the architectures mentioned above are inherently three-dimensional which require multi-step fabrication. Planar structures, on the other hand, can simplify this process tremendously, providing high integration densities at low cost. The advantages brought about by planar devices were touched on in Chapter 2, including a discussion of the planar SSD. Previously, SSDs have shown memory retention capabilities of ten hours at cryogenic temperatures, and minutes at room temperature [52]. Conduction was fundamentally controlled by using the charging and discharging of surface states in the channel. Later in this chapter, a novel planar structure based on the SSD, is proposed to demonstrate memory effects.

### 5.2 Conventional memory devices

The most important device for MOS memory technology today is the metal–oxide semiconductor (MOS) field-effect transistor (FET) [19, 20]. A simplified cross section of an n-channel MOS transistor is shown in Fig. 5.1. It consists of an insulating dielectric (usually, a thermally grown silicon dioxide film) sandwiched between a p-type silicon substrate with a surplus of holes and a top gate electrode made of conducting material. The source and the drain, having an abundance of electrons are intentionally slightly overlapped with the gate by n-doped ionised regions. When a positive voltage is applied to the gate, an electric field that penetrates through the insulator is set up, attracting electrons toward the surface of
5.2 Conventional memory devices

the substrate. An inversion layer designated as the channel is formed, allowing current to flow between the source and the drain regions. The two most applied technologies in the nonvolatile and volatile environments are flash and DRAM memories respectively. Their fundamental working principles and limitations are discussed separately below.

![Schematic cross-section representation of an n-channel MOS transistor structure](image)

**Figure 5.1** Schematic cross-section representation of an n-channel MOS transistor structure [19].

### 5.2.1 Flash memory

Since the first proposal of a floating gate device in 1967 [206], progression of memory technology has been intensively quick using its concept as the building block of many nonvolatile memory systems. The earliest nonvolatile device was the Erasable Programmable Read-Only Memory (EPROM). It was a floating gate based logic device similar to today’s flash memory but suffered from the fact that it could not be erased electrically. In fact, it required 20 minutes of UV-exposure to neutralise charge and erase a memory state [207]. This shortcoming was soon overcome with the Electrically Erasable Programmable Read-Only Memory (EEPROM) based on which the first generation of today’s widely dominant flash memory was invented by Dr. Fujio Masuoka of Toshiba in 1984 [206]. Today, over 90 % of nonvolatile memory production is based on flash [201].

Flash memory technology rests fundamentally on a floating gate capable of storing charge permanently, because it is completely surrounded by a high-quality insulator, thereby providing long-term memory. Complementary metal-oxide-
5.2 Conventional memory devices

Semiconductor (CMOS) field-effect transistors form the backbone of these devices; having a polysilicon floating gate surrounded by silicon dioxide ($\text{SiO}_2$), Fig. 5.2.

![Figure 5.2 A generic CMOS flash memory transistor [17].](image1)

Figure 5.3 (a) Read and write/erase operation principles of a flash memory. The schematic of the memory transistor illustrates the bias conditions that are employed for the case of CHE injection or F–N substrate electron injection write modes, and for the case of source-side electron extraction erase mode. (b) Circuit schematic of a one-transistor flash memory cell in a memory configuration (two memory cells are shown) [17].

![Figure 5.3](image2)

During programming, electrons (for the case of an $n$-channel device) from the silicon conduction band are injected through the oxide between the floating gate and the channel, also known as the tunnelling oxide, and subsequently stored on the floating gate. The stored electrons in the floating gate screen the mobile charge in the underlying channel, thus inducing a change in the conductivity of the channel,
5.2 Conventional memory devices

allowing the transistor threshold voltage to be electrically altered between a low and a high value, conventionally defined as the ‘1’ or erased state and the ‘0’ or written state (Fig. 5.3a). The threshold voltage shift $\Delta V_{TH}$ caused by the storage of the charge $Q_{FG}$ is given by $\Delta V_{TH} = -Q_{FG}/C_{CG}$ [207]. $C_{CG}$ is the capacitance between the control and the floating gate, and is given by $C_{CG} = \varepsilon A/t$, where $A$ is the capacitor area, and $\varepsilon$ and $t$ are the dielectric constant and thickness of the control dielectric respectively. In the memory-array configuration shown in Fig. 5.3b, each floating gate is accessed through a transistor connected to an array of word lines and bit lines. A circuit called a sense amplifier (S.A.) is used to distinguish whether the memory cell is storing 0 or 1 information (read operation) by sensing the current through the bit line, while accessing the cell is achieved by applying a read voltage $V_{\text{read}}$, at the gate or word line with a value between the two possible threshold voltages.

An erase operation involves removing electrons from the floating gate, hence returning $V_{TH}$ to a low state. Flash memories overcome the non-selective erasure that occurs in EEPROMs by means of lower cost large-scale integration-memory circuits. Once programmed, the cell is guaranteed to retain information for at least ten years (the industrial standard for nonvolatile memories), either under operation or with power turned off. Information can be lost if electron leak occurs, mainly through the tunnelling oxide and control dielectric. To maintain nonvolatility, relatively thick (about 10 nm) tunnelling oxides are required, a feature that causes difficulty in charging the floating gate as a large amount of energy will be needed for electrons to move into or from it.

Write/Erase mechanism

The physical mechanisms that are commonly used for this charge transfer are the field induced (Fowler–Nordheim, F–N) tunnelling [208] or channel hot-electron injection (CHE) [209] for the write operation and F–N tunnelling for the erase operation. F–N tunnelling is the flow of electrons through the energy barrier of the tunnelling oxide that has been modified by a high electric field for effective narrowing of the oxide barrier width. During F–N programming, a high voltage applied to the control gate injects electrons from the source, drain, or substrate (depending on the voltages at respective terminals) into the floating gate. The schematic description of the F-N electron injection is shown in Fig. 5.3a. During
5.2 Conventional memory devices

CHE injection, a large drain bias $V_D$ generates electrons with high energy at the drain side of the channel, while a large control gate $V_G$ bias, with $V_G > V_D$, generates an oxide field that favours the injection of the heated electrons over the Si–SiO$_2$ energy barrier.

The erase process can be performed by applying a high voltage to the source, drain, or to the substrate with the control gate grounded. It can also be accomplished by applying a negative voltage to the control gate with the other terminals connected at different positive voltages. Depending on the programming mechanism, flash memory arrays have two different configurations: NAND and NOR. NAND cells use F–N tunnelling to program the floating gate, while NOR cells use CHE injection for write operation. Both NAND and NOR cells use F–N tunnelling to erase the cell.

Limitations

The generation of highly energetic carriers during both programming operations introduces permanent damage, referred to as degradation, due to charge trapping in the tunnelling oxide. As a result, the flash memory can tolerate a limited number, typically $10^5$, of write/erase cycles (called its endurance). The charging times are slow, resulting in typical write times in the microsecond and millisecond ranges, respectively because the currents flowing through the tunnelling oxide are small.

Although flash memories have been designed as a solution to the scaling problem of conventional EEPROM devices, aggressive scaling of the transistor dimensions and the dramatic increase in the memory array size require a lower voltage memory cell design for the future.

Since the technology for the CHE injection or F–N tunnelling process has to support relatively high voltages (e.g., 8–9 V minimum and >12 V for CHE and F–N tunnelling programming, respectively), scaling of the flash memory cell is severely limited. This in turn impedes reduction in operating voltages. Constraints in scaling are primarily due to requirements of extremely low dielectric leakage (less than $\sim 10^{-14}$ A/cm$^2$) to ensure a ten-year data retention time. In the case of the control dielectric, an Oxide Nitride Oxide (ONO) composite dielectric is typically used, and allows for thinner control layers compared to silicon oxide layers, with much lower leakage currents. In the case of the tunnelling oxide, the requirements are even more stringent as it must be thin enough to allow a fast write/erase speed at reasonable
5.2 Conventional memory devices

voltage levels with negligible degradation after $10^5$ programming cycles, and thick enough to avoid charge loss during read or normal operations. Thus all scaling issues pertinent to flash memories are ultimately related to the reliability of the tunnelling oxide. In theory, to ensure ten-year data retention time, the tunnelling oxide could be scaled until electron flow through the full oxide thickness becomes significant (i.e., down to ca. 5 nm). However, stress-induced leakage current (SILC) \cite{210, 211} imposes a more stringent limitation on the tunnelling oxide thickness as a single point of high leakage in the oxide can discharge the conducting polysilicon floating gate layer. This explains why the tunnelling oxide of flash memories was set as thin as about 10 nm from the beginning, and has scarcely been thinned over five successive generations to its present limit of 7–8 nm. Consequently, the dimensions of the floating gate transistor have not been scaled as aggressively as those of the logic transistor, and therefore, the flash memory performance in terms of access time, write/erase speeds, and operating voltages has more or less stagnated. The increase in flash memory capacity (thus, the decrease in the cost/Mbit) has been mainly achieved through various array architectures such as the NAND structure \cite{212}, which reduces the cell size by connecting the cells in series. The multilevel charge storage (MLCS) approach, where more than one bit is stored inside a single memory transistor \cite{213} is another solution for higher bit storage density without the need for aggressive technology scaling.

5.2.2 Dynamic Random Access Memory (DRAM)

DRAM is commonly used as a main memory for personal computers due to their low cost and high-density capability. The benefit of DRAM is its structural simplicity: only one transistor and a capacitor (1T-1C) are required per bit (Fig. 5.4a), compared to four or six transistors in Static Random Access Memory (SRAM) \cite{214}. This allows DRAM to reach extremely high densities. The transistors and capacitors used are particularly small; billions can fit on a single memory chip.

Each bit of data is stored in a separate capacitor within an integrated circuit in the form of charge $Q_c$. By assuming that the capacitor’s common node is biased at
5.2 Conventional memory devices

$V_{CC}/2$, a logic 0 in the cell requires a capacitor with a voltage of $-V_{CC}/2$ across it, while a logic 1 in the cell requires a capacitor with a voltage of $+V_{CC}/2$ [17].

![Figure 5.4](image)

Figure 5.4 Schematic of (a) conventional planar one-transistor one capacitor DRAM cell, and (b) DRAM circuit in a memory-array configuration (two memory cells are shown) [6, 17].

Therefore, the charge stored in the capacitor is $Q = C_s(-V_{CC}/2)$ for logic 0 and $Q = C_s(V_{CC}/2)$ for logic 1, where $C_s$ is the capacitance of the storage capacitor. In the memory-array configuration, shown in Fig. 5.4b, each capacitor is accessed through a transistor connected to an array of word lines and bit lines. The sense amplifier (S.A.) distinguishes whether the memory cell is storing 0 or 1 information. The stored information is lost during reading and thus, a write cycle must always follow. The stored charge can be slowly depleted even during normal operation due to several leakage mechanisms unavoidably present, such as $p$–$n$ junction leakage, subthreshold conduction in the pass transistor, leakage through the capacitor dielectric, or through high-energy particle events (alpha particles, neutrons). Therefore, the refresh operation is integral to the proper operation of a DRAM.

During refresh operation, the cell content is read, and the data bit is overwritten, thus restoring the capacitor voltage to its proper value. Power consumption is increased during refreshing operation, and imposes a serious constraint in transistor scaling. In order to reach a long refresh time, the leakage of the access transistor must be low, needing relatively high threshold voltages ($> 0.5$ V). This is the reason why DRAM transistors have not scaled as fast as logic transistors, and generally have longer channel lengths, thicker gate oxides, and higher operating voltages. Another constraint in DRAM cell scaling is imposed by the scaling of the storage
5.2 Conventional memory devices

capacitor. The voltage available to the sense amplifier for signal detection is proportional to the $C_S/C_B$ ratio, therefore, the storage capacitance $C_S$ needs to be as large as possible in order to create a large enough signal (≈100 mV) on the bit line. $C_B$ is an inevitable parasitic bit line capacitance coming from the transistor junction capacitance and bit-line wire capacitance. $C_S$ also must significantly remain large enough to reduce the sensitivity of the cell to soft errors that originate from high-energy particle events. Typical $C_S$ values are in the range of 25 – 30 fF, and these values have been constant over many generations of DRAM technology. According to the equation for the memory cell capacitance, $C_S = \varepsilon\varepsilon_o A/d$, where $\varepsilon$ is the relative dielectric constant of the capacitor film material, $\varepsilon_o$ is the dielectric constant of free space, $A$ is the capacitor area, and $d$ is the thickness of the capacitor film, the approaches that have been used in order to maintain a constant $C_S$ are as follows. Firstly, in the simple planar cell in Fig. 5.4a, the pure oxide capacitor film with a dielectric constant of 4 is replaced by a multilayer oxide/nitride/oxide (ONO) or nitride/oxide (NO) film with a dielectric constant of 7, followed by capacitor film thickness $d$ scaling. Thereafter, three-dimensional (3D) capacitor structures (stack or trench) are established in place of the conventional planar cell to raise the capacitance per unit area.

It is expected that future DRAM capacitors will need dielectric materials having high $\varepsilon$ values such as Tantalum Pentoxide ($Ta_2O_5$, $\varepsilon = 25$) and, ultimately, materials with even higher $\varepsilon$ values such as Barium Strontium Titanate ($BST$, $\varepsilon = 500$) [215]. Although these exotic materials would considerably increase the capacitance per unit area, and thus the device integration density, they are not at this time compatible with the existing DRAM manufacturing process. All in all, it is indeterminate whether the DRAM memory cell will scale at gigabit densities due to the innate constraints of low-leakage access transistors and the need for large storage capacitance.

5.2.3 New developments

Memory device mechanics have met their scaling limit owing to reasons unique to their mechanisms as discussed above. To overcome these technological constraints, new memory concepts are needed. Various memory alternatives actively
5.2 Conventional memory devices

investigated include the phase-change RAM (PCRAM) which is a type of non-volatile random-access memory. PCRAMs exploit the unique behavior of using heat to switch between amorphous and crystalline states in a number of new materials, e.g. chalcogenide glass compounds such as Silver Indium Antimony Tellurium (AgInSbTe) and Germanium Antimony Tellurium (Ge$_2$Sb$_2$Te$_5$ or GST). The ability to achieve a number of distinct intermediary states in this type of memory gives the ability to hold multiple bits in a single cell. Newer PCRAM technology has been trending in a couple of different directions. Some groups have been directing a lot of research towards attempting to find viable material alternatives to GST, with mixed success, while others have developed the idea of using a GeTe - Sb$_2$Te$_3$ superlattice in order to achieve non thermal phase changes by simply changing the coordination state of the Germanium atoms with a laser pulse, and this new Interfacial phase-change memory (IPCM) has had many successes and continues to be the site of much active research [216]. Known for being non-volatile and scalable, PCRAMs have been touted as a promising solution that combines the advantages of DRAM and flash. The cost factor, however, is not expected to be lower and the relatively large reset current may make it applicable to certain applications but would definitely lose its attraction on the consumer electronics market.

![Floating gate nonvolatile memory; and (b) a nanocrystal nonvolatile memory](image)

**Figure 5.5** The structures of (a) a floating gate nonvolatile memory; and (b) a nanocrystal nonvolatile memory [218].

Another promising alternative for low-cost ultradense data storage comes from the use of modified flash memory structures, e.g. nanocrystal memories [217]. The
5.3 Planar nano-memory devices

storage of charges in a layer made of discrete nanocrystals replaces the conventional floating gate (Fig. 5.5) and is considered to be a promising candidate for the next generation of nonvolatile memories due to its high operation speed, good scalability, and superior reliability. At the same time, it possesses the potential to perform fast DRAM programme/erase speeds and long flash memory retention times simultaneously [218]. Again, the dual memory technology (volatile and nonvolatile) is a promising feature of nanocrystal memories that suits the present and future portable electronics market. However, acquiring uniform layers of crystals repetitively is still a struggle [219] and the charge leakage is still dependent on the layer thickness and surrounding oxide passivation. Alternative suggestions, such as using very thin high-k dielectric films like AlN [220], have been made but these will also ultimately reach scaling limits given their vertically layered architecture.

For this reason, recent literature for memories have focussed on resistive switching in planar two-terminal devices [221-225] that basically show resistive, capacitive, and/or inductive properties that are hysteretic when subject to time-dependent perturbations. Switching between a high and low state is achieved by means of an appropriate electrical pulse, and the state can subsequently be read out at low bias. The developments in this area are also driven by emerging large-area microelectronic applications, such as rollable displays [23], electronic paper [24], contactless identification transponders [25, 26], and smart labels [27]. Thus, to overcome the scalability issue and quickly meet consumer driven market demands, new architectures would be necessary to deliver DRAM and flash memories. It must be drop-in compatible with existing memory and have favorable cost, power, reliability, and performance characteristics; and planar devices may provide the ultimate solution.

5.3 Planar nano-memory devices

The numerous advantages derived from planar architectures in nanoelectronics were comprehensively covered in Chapter 2, the main benefits being simplified fabrication with single-step lithography and multiplication using arrays without incurring high costs. It is envisioned that these advantages profit memory devices as well, possibly overcoming some of the issues conventional memory devices have.
5.3 Planar nano-memory devices

Planar nanoelectronic memories using capacitively coupled side-gated transistors (SGTs) have already taken shape in the form of a set-reset (SR) latch shown in Fig. 5.6 [226]. Fabricated on high mobility InGaAs/InP modulation doped 2DEG material, the source of each device is connected to the gate of the other.

![Atomic force microscope image of novel SR-latch proposed by Sun et al. [226]. Right: Truth table for SR latch.](image)

Although room temperature operation has been achieved, the device is somewhat limited in applications because the optimal gating efficiency requires voltage shifts of 1.5 V to the drains of each SGT, effectively so due to the non-zero threshold of the transistor. SSDs (already covered in detail in Chapter 2), on the other hand, provide tuneable threshold, to even zero volt.

Besides displaying rectifying behaviour, SSDs have also demonstrated memory effects [52]. The IV curves showed distinguishable hysteresis (Fig. 5.7a), when a large enough reverse or forward bias was applied, with time scales in the order of minutes at room temperature and hours at cryogenic conditions. The hysteresis was proposed to be a result of charging and discharging of surface states along the nanochannel (Fig.5.7b) and was later verified by means of a Monte Carlo interpretation [146]. Such memory devices consisting of just two terminals, also termed memristors, though seemingly sufficient, leave a lot to be desired in nonvolatile memory applications. This is because only one terminal can be used for both reading and writing, thus incorporation into large network of memory cells would require new memory architecture in circuitry. Hence, a three terminal extension of the SSD would seem more appropriate to exhibit DRAM or flash memory.
5.3 *Planar nano-memory devices*

A novel device (Fig. 5.8), built on the functionality of the SSD introduces a memory storage region nearby the nanochannel, where self-terminating encapsulation of charge is made possible with the presence of two nanogaps. Owing to their dimensions and surface states present on the walls of the nanogaps, charging and discharging of the surface states can alter the effective gap sizes thus controlling

![Figure 5.7](image)

**Figure 5.7** (a) Low temperature I-V characteristics of a self-switching memory. (b) Illustration of surface state charging and discharging leading to SSD memory effect. When the reverse bias exceeds $V_{th^{-}}$, the surface states discharge by charge transfer into the channel (reducing the surface depletion region). Conversely, when the forward bias exceeds $V_{th^{+}}$, the surface states are recharged (increasing the surface depletion region) [52].

![Figure 5.8](image)

**Figure 5.8** The proposed memory device structure. The conduction band profile along the dotted line is illustrated in Fig. 5.9.
5.3 Planar nano-memory devices

electron transfer, leading to interesting memory effects. The notion of interface-charge which is central to the memory effect in this device will be presented in the next section. It is important to note that, practically, the interface-charge density is a dynamic quantity that varies with electron/current flow. However, to extract valuable data with respect to the variables under study, the interface-charge density is kept constant in the simulations performed later.

In the absence of charge in the storage region, current flow is high in the nanochannel (low resistance), while the opposite effect occurs when it is charged (high resistance). Thus the I-V characteristics are influenced, having different threshold voltages, for the charged (off) and discharged (on) states. Conventional ‘0’ and ‘1’ bit states can also be assigned respectively. The operation is similar to the floating gate of a flash memory, but needs no isolating oxide around the gate or hot electron effects for charging and discharging. However, the memory retention capability is not expected to reach nonvolatile standards, at least not at this initial stage of the device development. It is further envisaged that a refresh operation akin to DRAM will be required at the end of each reading cycle to restore the charging capacity. Therefore the device may identify better with the RAM and will be referred to as the planar RAM (pRAM) henceforth.

The memory region can be charged by applying a large enough negative bias at the memory gate, drawing electrons into the storage region. An illustration of the charging effects across interfaces is shown in Fig. 5.9 for equilibrium and charged states (The profile line is shown in Fig. 5.8).

In the charged state, the memory charge causes the depletion region to prolong further into the channel, inhibiting current, compared to the equilibrium state. Subsequently, electron depletion is enhanced around the nanogaps, thereby self-terminating the charging effect by the pinched off gaps and retaining charge. Conversely, discharging can be done by applying a large positive bias at the memory gate. In this way, a binary memory system with either a full or empty memory region is achieved.

The pRAM has advantages over other types of memory devices because the charge-trapping is controllable and tuneable by varying its size, aspect ratio to the channel and gap sizes. In addition, a single high resolution lithography step enables fabrication of large memory array structures, reducing cost significantly. Operating
5.3 Planar nano-memory devices

voltages are also scaled down with the eradication of vertical architecture and by zero threshold. Thus, the device can potentially be applied to novel device application areas, such as organic, flexible, and printed electronic devices [204].

![Diagram of planar nano-memory devices](image)

Figure 5.9 The conduction band schematic in the pRAM along the dotted line shown in Fig. 5.8. Memory charge extends the depletion region in the channel inhibiting current.

5.3.1 Surface states and space charge

The principle of charge encapsulation in the pRAM is exclusively based on the surface states and surface charge in the semiconductor-insulator interfaces; the theory of which is described in this section. When chemical bonds are broken by cleavage of a semiconductor surface, dangling bonds leads to space-charge effect. The distribution of surface states at the n-type non-degenerate semiconductor interface causes upward band bending. The schematic shown is shown in Fig. 5.10 where it is assumed the surface state is half-filled before equilibrium. $E_F$ (bulk) is closer to the conduction band, which is higher than $E_F$ (surf) under disequilibrium. The electrons will transfer from the bulk to the surface. $E_F$ (bulk) drops and $E_F$ (surf) rises until equilibrium is achieved. At equilibrium, the energy bands bend upward in the semiconductor towards the surface (Fig. 5.10b).
5.3 Planar nano-memory devices

The band bending is due to the poor screening of the surface charge-induced electric field by the low concentration of free carrier density in the semiconductor. In metals, the free electron density is $\sim 10^{22} \text{ cm}^{-3}$ with a short screening length on the order of atomic sizes, while the screening length in a semiconductor is in the order of $\sim 100 \text{ Å}$ with a free carrier density of $\sim 10^{17} \text{ cm}^{-3}$, which creates a space charge region near the semiconductor surface. Physical properties, such as energy band structure, free carrier density, and local conductivity, will also be changed in the space charge region compared with the bulk [227].

![Schematic electron energy levels near the surface of a clean semiconductor: (a) disequilibrium and (b) equilibrium](image)

**Figure 5.10** Schematic electron energy levels near the surface of a clean semiconductor: (a) disequilibrium and (b) equilibrium [230].

Figure 5.11 summarises three kinds of space charge regions schematically [227-229]. Only the n-type semiconductor is shown where the majority carriers are electrons, that is, the density of free electron carriers ($n_e$) is higher than that of hole carriers ($n_h$). In the upward band bending condition shown in Fig. 5.11c, negative charges exist at the surface and positive charges accumulate near the surface, causing a decrease of $n_e$ and an increase of $n_h$. This space charge region is called the depletion layer. At thermal equilibrium, the space charge is balanced by a net charge in surface states, therefore, negative values of space charge indicate upward bending, increasing the distance from the Fermi Level to the bottom of the conduction band. In one-
5.3 Planar nano-memory devices

dimensional space, the electric potential \((V)\) of a point \((z)\) relative to the bulk of the semiconductor can be described by Poisson’s equation,

\[
\frac{\partial^2 V}{\partial z^2} = \frac{-q(z)}{\varepsilon_r \varepsilon_0}
\]  

(5.1)

where \(q\) is the space charge density and \(\varepsilon_r\) and \(\varepsilon_0\) are the relative dielectric constants of the semiconductor and the vacuum permittivity, respectively.

---

**Figure 5.11** Schematic diagrams showing the energy levels and free charge carrier densities (logarithmic scale) from the n-type semiconductor surface to the bulk. The blue dotted lines indicate the corresponding space charge region of thickness, \(D\). \(n_e\) = free electron density; \(n_h\) = free hole density; \(n_i\) = intrinsic carrier density [227].
5.3 Planar nano-memory devices

Figure 5.12 (a) GaAs substrate with 200 nm width insulating trenches (air) with a 100 nm gap between the trenches. (b) Conduction band diagram along the dotted line showing the magnitude of band bending for different interface-charge densities.

To realise the conduction band bending along a semiconductor-insulator interface such as GaAs-Air, the SILVACO simulation software is used to describe the effect of introducing different values for interface-charge density (Fig. 5.12). The carrier density in GaAs is fixed at $1 \times 10^{17}$ cm$^{-3}$. The results follow Eq. (5.1) and the band bending heights increase proportionately to the value of interface-charge densities. Band bending can also be treated as a potential barrier that dissipates quadratically with distance ($z$) from the semiconductor interface. Therefore, the potential and the related behaviour can be approximated using the formalism of the widely understood Schottky barrier [231] which is presented next.

5.3.2 Current transport in Schottky barriers

When a metal is brought into contact with a semiconductor (Fig. 5.13b), Fermi levels tend to line up at thermal equilibrium (Fig. 5.13c) causing the formation of a potential barrier, known as Schottky barrier, which determines the electron transport through the interface. The barrier height which is critical for current transport, is independent of the metal work function and relies completely on the semiconductor doping and its surfaces properties, for example density of surface states [20].
Two mechanisms dominate current transport in Schottky barriers, the first being thermionic emission, which refers to the transfer of electrons over the potential barrier, and the second being quantum mechanical tunnelling through the barrier [20]. The first mechanism, thermionic emission, dominates at room temperature in moderately doped semiconductors under forward bias and makes two assumptions: first, that the barrier height is much greater than $kT \phi_b / kT$, and second, that collisions within the depletion region is neglected. The current flow is then described by equations that follow.

The current density from the semiconductor to the metal can be described as

$$ J_{S \rightarrow M} = A T^2 \exp \left( \frac{-q \phi_b}{kT} \right) \exp \left( \frac{qV}{kT} \right) $$

(5.2)
where \( T \) is temperature (=300K), \( \phi_b \) is the barrier height, \( q \) is the electron charge, \( k \) is the Boltzmann’s constant and \( A^* \), known as the Richardson’s Constant, is

\[
A^* = \frac{4\pi q m^* k^2}{h^3}
\]  

(5.3)

At thermal equilibrium, the barrier height is the same for electrons moving from the metal to the semiconductor and vice versa, therefore \( V = 0 \), and the current density in the opposite direction is

\[
J_{M\rightarrow S} = A^* T^2 \exp\left(\frac{-q\phi_b}{kT}\right)
\]  

(5.4)

By adding (5.2) and (5.4), the total current density and the net current flow can be obtained:

\[
J_T = A^* T^2 \exp\left(\frac{-q\phi_b}{kT}\right) \left\{ \exp\left(\frac{qV}{kT}\right) - 1 \right\}
\]  

(5.5)

\[
I = S A^* T^2 \exp\left(\frac{-q\phi_b}{kT}\right) \left\{ \exp\left(\frac{qV}{kT}\right) - 1 \right\}
\]  

(5.6)

where \( S \) is the area of Schottky contact. Typical Richardson’s constants for GaAs and silicon are \( 8.6\times10^4 \) and \( 2.6\times10^6 \) A/m\(^2\)/K respectively.

Equation (5.6) plotted for a range of Schottky barrier heights for a silicon/metal interface at \( T = 300K \) is shown in Fig. 5.14.

From the graph, an idea of the thermionic emission expected for different potentials can be conceived. For example, when a potential of \( V = 0.2 \) eV exists between the two regions, and for a barrier height of \( \phi_b = 1.1 \) eV, current equivalent to 1 electron per second is expected. This is applicable to the pRAM in which the barrier heights imposed by the nanogaps determine memory retention times. Assuming memory charging doubles the equilibrium carrier concentration of \( 10^{17} \) cm\(^{-3}\), this level of current by thermionic emission will equate to charge retention times in the order of a day at room temperature.
5.3 Planar nano-memory devices

Figure 5.14 Current transport for a range of Schottky barrier heights using Eq. (5.6) for a contact area of depth = 200 nm and width = 50 nm of a Silicon substrate.

From the graph, an idea of the thermionic emission expected for different potentials can be conceived. For example, when a potential of \( V = 0.2 \) eV exists between the two regions, and for a barrier height of \( \phi_b = 1.1 \) eV, current equivalent to 1 electron per second is expected. This is applicable to the pRAM in which the barrier heights imposed by the nanogaps determine memory retention times. Assuming memory charging doubles the equilibrium carrier concentration of \( 10^{17} \) cm\(^{-3} \), this level of current by thermionic emission will equate to charge retention times in the order of a day at room temperature.

Simulated plots of the conduction band energies for different gap sizes between insulation trenches in Fig. 5.15 reveal barrier heights of 1 eV are attainable for a gap size of 40 nm and below. This is in the region of the required barrier height for the floating gate principle, however, the advantage of achieving barrier heights using nanogaps is that they are independent of work function and can be controlled by bias, when compared to conventional memory devices.
5.4 Simulation results

Physical modelling is an invaluable tool when a device or semiconductor structure is completely new and has no history. For the novel pRAM, analytical studies can be performed from which critical memory parameters and I-V characteristics can be extracted. Most notably device design will be related to charge retention time, writing time and operational voltages.

SILVACO has built-in specific features for modelling of conventional flash memories; these include models for Fowler-Nordheim tunnelling [232], and Hot Electron injection (also known as Lucky Electron Model) first proposed by Tams et. al. Both these models are for carrier transport across the insulating gate oxide, which are not relevant in the pRAM where carrier transport is essentially through depleted semiconductor regions/gaps for sufficiently low electric fields. Device physics introduced in Chapter 3 is used where the transport equations dictate the relationship between two or more quantities, through physical constants and basic material/device parameters such as dielectric constant, geometry, conductivity, etc. The mobility, carrier statistics and recombination models are also appropriately assigned to develop the physical modelling program.

Both steady state and transient simulations are performed to fully characterise the pRAM. The steady state results provide very useful information on the threshold shift possibilities in a given geometry; and the transient simulations are used to
5.4 Simulation results

derive estimates for critical parameters such as charge writing time, memory retention time and operational voltages. These fundamentally govern the memory volatility of the device and write speeds required to meet modern application requirements.

5.4.1 Material, models and parameters

Modern silicon-on-insulator (SOI) technology is the focal point of today’s memory devices. SOI is a specially made substrate with a high quality doped single crystalline silicon layer, of down to 200 nm, on top of a ~100 nm thin insulating Buried Oxide (BOX) layer (silicon dioxide) [133]. Similarly, a two-dimensional n-type silicon layer is adopted in this work, closely related to previously reported planar device physical modelling using SOI [133]. A thickness of 200 nm and n-type doping of 1×10^{17} \text{ cm}^{-3} (sheet concentration of 2×10^{12} \text{ cm}^{-2}) are assigned to the silicon layer. The insulating trenches are filled with air although oxide-filled trenches can bring about increased gate capacitance and potentially reduce operation voltages and even increase memory retention time. This is essentially because oxide growth in the narrow trench widths ~200 nm, is practically not feasible due to the non-conformal surface impinging nature of oxide growth [233]. Despite the challenge, high quality oxide filling of trenches with aspect ratio up to 6 was achieved [205]. This has significant implication to the pRAM operation and is worthwhile to consider at the fabrication stage. A simulation for the pRAM containing oxide-filled trenches will be performed to provide supplementary information at the end of the chapter. The device dimensions (in microns, not drawn to scale) are shown in Fig. 5.16. The channel width W and gap sizes G1 and G2 are varied to decide on optimal dimensions. The interface-charge density (σ) at the semiconductor-air interfaces is set to be -1×10^{12} \text{ cm}^{-2} [20, 133]. Finally, the simulations are performed at room temperature conditions (300K).

Under model specifications, the drift-diffusion transport equations are deployed to solve potential and electron/hole concentrations isothermally. The mobility is modelled as doping concentration dependent by invoking the CONMOB syntax in the mobility statement, along with the Shockley Read Hall carrier generation and recombination model [172].
5.4 Simulation results

5.4.2 Steady-state characteristics

Output characteristics

The output characteristic ($I_D$ vs $V_D$) follows the form of a typical SSD as expected (Fig. 5.17). It is controlled by the channel width and the interface-charge density ($\sigma$). Since $\sigma$ is fixed, a suitable channel width can be selected where the channel will be close to pinched-off for zero gate bias. This allows for low operational current and potentially a high on/off ratio. From the transfer characteristics below, the on/off ratios are determined which will provide the choice of channel width to select.

![Diagram of pRAM device](image)

**Figure 5.16** The geometric dimensions of the simulated pRAM device. (in microns, not drawn to scale).

Transfer characteristics

The gating effect on different channel widths is studied by sweeping the gate bias from -10 to 10 V (Fig. 5.18a). The highest on/off ratio is obtained from $W = 180$ nm making it suitable for controlling conduction current in the channel. With the memory region charged ($V_{G2} = -10$V), the on-off ratio remains high for $W = 180$ nm (Fig. 5.18b), demonstrating that this dimension is reliable for a large range of biasing (as may be achieved using two gates). The availability of two gates is an added advantage of the pRAM as it may suit some applications without needing to switch to a new device or alter configurations to the circuit.
5.4 Simulation results

![Figure 5.17 I-V characteristics for the pRAM for different channel widths varied from 180 to 230 nm. Both gates were fixed at 0 V.](image)

5.4.3 Transient characteristics

Based on the device having a channel width of 180 nm, transient simulations are performed in this section to gain information on the memory retention times. The values obtained are used to improve the design in a continuous process where the objective is to enhance the memory time significantly. The two gaps which control the barrier heights play a major role in this process and will be studied first. Following this, the dependence on the pulse width of the memory bias will be considered. The memory bias is initially set to ±10 V but the effect of increasing or decreasing this magnitude will be reported. This will ultimately set the optimal write/erase pulse time and voltage for a given structure. Generally, the charge storing pulse (low current) is the write pulse where information is stored and the discharge pulse is the erase pulse. However, this device is new and would be imperative to fully characterise it. Therefore, memory retention times will be observed for both pulses to understand their respective charge storage capabilities and leakage currents. Finally, a set of alternative designs such as creating additional memory regions (array) and oxide-filled trenches will be investigated.
5.4 Simulation results

Figure 5.18 Transfer characteristics for the pRAM for different channel widths varied from 160 to 200 nm, with the memory gate $V_{G2}$ held at (a) zero bias and (b) negative bias. The 180 nm wide channel gives the highest on/off ratio.

The transient simulation is implemented by applying two bias pulses, negative and positive, on the memory electrode which is otherwise at equilibrium, as shown in Fig. 5.19b. The pulse width is also nominally selected to be $10^{-3}$ s. A constant low bias of $V_D = 1$ V is applied at the drain to keep the channel at pinched-off mode. The resultant drain current is plotted against time. Fig. 5.19c identifies the important values to be extracted, namely the on and off currents and their respective charged and discharged memory retention times. Memory time is taken to be up to the point where $I_D$ recovers exponentially to equilibrium.
5.4 Simulation results

Figure 5.19 (a) The memory structure showing bias conditions for simulations, (b) the bias signals applied to the memory gate and (c) a typical transient response of the drain current from which the key parameters are extracted.

Gap size dependence

The study performed in this section is of utmost importance to the working operation of the pRAM as discussed in 5.3. A thorough analysis is provided for gap size dependence. In order to establish a region of gap size to work with, the two gaps are initially set at a 1:1 ratio, beginning with 20 nm ranging up to 120 nm. Other parameters are as per Fig. 5.19a. Using the results, different G1 to G2 ratios will be studied subsequently.

The transient result for G1 = G2 = 60 nm is shown in Fig. 5.20. Information extracted from the transient drain currents for the six sets of gaps is given in Table 5.1. The dashed non-values represent absence of memory time; i.e. the current drops
off exponentially after 3s. This data is plotted in Fig. 5.21. The best set of results is achieved with the 60 nm gap sizes (circled).

Figure 5.20 Drain current transient response for the pRAM with equal dimensions for both gaps (60 nm). The pulse magnitude is 10 V and the pulse width is $10^{-3}$ s. Right: Zoomed current response after the positive pulse.

<table>
<thead>
<tr>
<th>G1 (nm)</th>
<th>G2 (nm)</th>
<th>Off current ($10^{-14}$ A)</th>
<th>On current ($10^{-7}$ A)</th>
<th>On/Off ratio ($\times10^6$)</th>
<th>t1 (ms)</th>
<th>t2 (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>20</td>
<td>0.8</td>
<td>0.05</td>
<td>0.7</td>
<td>380</td>
<td>-</td>
</tr>
<tr>
<td>40</td>
<td>40</td>
<td>1.0</td>
<td>8.7</td>
<td>8.7</td>
<td>200</td>
<td>-</td>
</tr>
<tr>
<td>60</td>
<td>60</td>
<td>2.1</td>
<td>1.8</td>
<td>8.6</td>
<td>150</td>
<td>10</td>
</tr>
<tr>
<td>80</td>
<td>80</td>
<td>2.4</td>
<td>1.2</td>
<td>4.8</td>
<td>120</td>
<td>2.4</td>
</tr>
<tr>
<td>100</td>
<td>100</td>
<td>3.0</td>
<td>1.0</td>
<td>3.3</td>
<td>90</td>
<td>1.2</td>
</tr>
</tbody>
</table>

Table 5.1 On/Off ratios and memory retention times for gap dimensions of a 1:1 ratio ranging from 20 to 100 nm.
5.4 Simulation results

Figure 5.21 On/Off ratios and memory retention times for gap dimensions of a 1:1 ratio ranging from 20 to 100 nm. A high on/off ratio along with relatively long $t_1$ and $t_2$ are achieved for $G_1 = G_2 = 60$ nm (circled).

The conduction band contour plots are shown in Fig. 5.22 for the two pulses in contrast to the equilibrium condition ($V_{G2} = 0$). The charged and discharged memory regions are clearly visible from the plots. The profile lines in Fig. 5.22d confirm the charging effects across interfaces that were discussed in Fig. 5.9 showing the possibility of barrier height modulation at the nanogap.

Following from above, it can be understood that different combinations of gap sizes may provide more effective encapsulation of charges in the storage region. Therefore, combinations of gap sizes between 20 and 80 nm are studied to establish a pair of gap sizes that yields a large on/off ratio and substantial memory times for both $t_1$ and $t_2$. The obtained results for these combinations are tabulated in Table 5.2 and also plotted in Figs. 5.23a and b.

The on/off ratios are in the $10^6$ range for 80% of the data points. As for the memory retention times, more analysis is required to establish how they are affected. Firstly, charged memory retention times ($t_1$) are not difficult to achieve for all the combinations. They are longest when $G_1$ and $G_2$ are as small as possible. This is observable from Table 5.3a where $t_1$ times are ranked. In fact, it is established that $t_1$ is proportional to the average barrier height per nanometer of the combined gap size,
combined $\mathcal{V}_b / (G1 + G2)$. This is because after the write pulse is removed, the electron rich memory region attempts to discharge back to equilibrium. If the barrier heights of the two gaps are high, the charges are retained within for longer. The values of these barriers are presented in Table 5.4.

**Figure 5.22** (a) Conduction band contour plots and (b) potential barrier heights along the dotted line (equilibrium, charged and discharged conditions). $G1 = G2 = 60$ nm. The conduction in the channel is controlled by the charging and discharging. The barrier height can be elevated at charged condition thus pinching off the nanogap.
5.4 Simulation results

Table 5.2 On/Off ratios and memory retention times for combinations of gap sizes.

<table>
<thead>
<tr>
<th>G1 (nm)</th>
<th>G2 (nm)</th>
<th>Off current ($\times 10^{-11}$ A)</th>
<th>On current ($\times 10^{-7}$ A)</th>
<th>G1 (nm)</th>
<th>G2 (nm)</th>
<th>Off current ($\times 10^{-11}$ A)</th>
<th>On current ($\times 10^{-7}$ A)</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>20</td>
<td>0.6</td>
<td>0.7</td>
<td>20</td>
<td>20</td>
<td>2.5</td>
<td>0.1</td>
</tr>
<tr>
<td>40</td>
<td>40</td>
<td>3.3</td>
<td>2.3</td>
<td>40</td>
<td>40</td>
<td>1.0</td>
<td>0.7</td>
</tr>
<tr>
<td>60</td>
<td>60</td>
<td>9.6</td>
<td>0.39</td>
<td>60</td>
<td>60</td>
<td>2.7</td>
<td>2.3</td>
</tr>
<tr>
<td>80</td>
<td>80</td>
<td>2.5</td>
<td>0.23</td>
<td>80</td>
<td>80</td>
<td>2.7</td>
<td>1.6</td>
</tr>
</tbody>
</table>

Figure 5.23 (a) On/Off ratios and (b) memory retention times for combinations of gap sizes. To achieve equally long retention times for both on and off pulses, G1 = 20 nm and G2 = 60 nm (circled) are selected as the gap sizes.
5.4 Simulation results

Table 5.3 Ranking of memory retention times (a) $t_1$ and (b) $t_2$.

<table>
<thead>
<tr>
<th>Rank</th>
<th>G1 /G2 (nm)</th>
<th>$t_1$ (s)</th>
<th>Rank</th>
<th>G1 /G2 (nm)</th>
<th>$t_1$ (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20/20</td>
<td>0.38</td>
<td>9</td>
<td>80/40</td>
<td>0.16</td>
</tr>
<tr>
<td>2</td>
<td>20/40</td>
<td>0.30</td>
<td>10</td>
<td>80/40</td>
<td>0.16</td>
</tr>
<tr>
<td>3</td>
<td>40/20</td>
<td>0.27</td>
<td>11</td>
<td>60/60</td>
<td>0.15</td>
</tr>
<tr>
<td>4</td>
<td>20/60</td>
<td>0.22</td>
<td>12</td>
<td>60/80</td>
<td>0.14</td>
</tr>
<tr>
<td>5</td>
<td>40/40</td>
<td>0.20</td>
<td>13</td>
<td>80/20</td>
<td>0.13</td>
</tr>
<tr>
<td>6</td>
<td>60/20</td>
<td>0.20</td>
<td>14</td>
<td>80/80</td>
<td>0.12</td>
</tr>
<tr>
<td>7</td>
<td>60/40</td>
<td>0.20</td>
<td>15</td>
<td>20/80</td>
<td>0.08</td>
</tr>
<tr>
<td>8</td>
<td>40/60</td>
<td>0.16</td>
<td>16</td>
<td>40/80</td>
<td>0.08</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rank</th>
<th>G1 /G2 (nm)</th>
<th>$t_2$ (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20/60</td>
<td>0.28</td>
</tr>
<tr>
<td>2</td>
<td>20/80</td>
<td>0.16</td>
</tr>
<tr>
<td>3</td>
<td>40/80</td>
<td>0.14</td>
</tr>
<tr>
<td>4</td>
<td>40/60</td>
<td>0.12</td>
</tr>
<tr>
<td>5</td>
<td>60/80</td>
<td>0.07</td>
</tr>
<tr>
<td>6</td>
<td>60/60</td>
<td>0.01</td>
</tr>
<tr>
<td>7</td>
<td>80/60</td>
<td>0.003</td>
</tr>
<tr>
<td>8</td>
<td>80/80</td>
<td>0.002</td>
</tr>
</tbody>
</table>

Table 5.4 Average barrier heights per nanometer of combined gap size relating to the dependency of $t_1$.

<table>
<thead>
<tr>
<th>G1 /G2 (nm)</th>
<th>Average barrier per nm of combined gap size (eV/nm)</th>
<th>$t_1$ (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>20/20</td>
<td>0.52</td>
<td>0.38</td>
</tr>
<tr>
<td>20/40</td>
<td>0.34</td>
<td>0.30</td>
</tr>
<tr>
<td>40/20</td>
<td>0.33</td>
<td>0.27</td>
</tr>
<tr>
<td>20/60</td>
<td>0.25</td>
<td>0.22</td>
</tr>
<tr>
<td>40/40</td>
<td>0.22</td>
<td>0.20</td>
</tr>
<tr>
<td>60/20</td>
<td>0.20</td>
<td>0.20</td>
</tr>
<tr>
<td>60/40</td>
<td>0.17</td>
<td>0.20</td>
</tr>
<tr>
<td>40/60</td>
<td>0.19</td>
<td>0.16</td>
</tr>
</tbody>
</table>

From the data in Table 5.2, $t_2$ looks much more difficult to attain. But its appearance is made possible by wider G2 (i.e. 60 and 80 nm) and enhanced by narrower G1. This can be related back to the barrier height elevation found in Fig. 5.22. Therefore, geometrical optimisation is necessary in tuning the barrier height such that desired performance is achieved, as there is always the trade-off of being able to push...
5.4 Simulation results

carriers through versus retaining them without leakage. Barrier height modulations also depend on higher bias or a longer pulse time; which will be considered later. The study on gap sizes was more complicated than it seemed as they functioned in tandem. However, it was dealt with carefully to investigate these dependencies and form a conclusion. t1 is much easier to achieve and a compromise is necessary to evoke t2. For example, based on the rankings of the two memory times in Table 5.3, G1 = 40 nm, G2 = 80 nm is the worst candidate for t1 but seems to do well for t2. In this respect, Fig. 5.23b is referred to in order to identify the dimensions that provide equally long times for both memories. This is the circled data for G1 = 20 nm and G2 = 60 nm where both memory times are approximately 0.3s. It is no surprise that this set of dimensions rendered the best memory times for reasons already discussed above. Figure 5.24 plots the transient response of this structure. The potentials in the two gaps as a function of time are also plotted in the same graph. The memory retention times, t1 and t2, are also the times taken for the potentials at the gaps to return to their equilibrium values.

Figure 5.24 Drain current transient response for the pRAM with G1= 20 nm and G2 = 60 nm, showing a pair of substantially long memory retention times when compared to other gap dimensions. The potentials in the two gaps demonstrate that the barrier heights in the two gaps determine the leakage rate.
5.4 Simulation results

Pulse width dependence

The pulse width (pw), which has been $10^{-3}$ s thus far is varied to investigate the dependency. Since this relates to the period of carrier injection into and out of the memory region, too short a pulse width will lead to under-storage of the memory region. It is expected that the pulse width can be increased up to a saturation where holding the pulse any longer will not inject more carriers; depending on carrier density and size of memory region. An optimal pulse width can thus be deduced. Table 5.5 provides the data and Fig. 5.25 shows the response with reference to $pw = 10^{-3}$ s. At first look, $pw = 10^{-1}$ s may seem to deliver better memory time, but the pulse width is substantially more and has to be subtracted for a fairer judgement. The resultant memory time is still lower or equal to the one with $pw = 10^{-3}$ s. The table confirms the predicted behaviour of pulse width variance, where beyond $pw = 10^{-3}$ s, the memory times cannot be improved. Therefore, the pulse width of $10^{-3}$ s can be regarded optimal for this structure.

<table>
<thead>
<tr>
<th>G1 = 20 nm</th>
<th>G2 = 60 nm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pulse width (s)</td>
<td>Off current ($\times10^{-14}$ A)</td>
</tr>
<tr>
<td>$10^{-5}$</td>
<td>4.5</td>
</tr>
<tr>
<td>$10^{-4}$</td>
<td>3.0</td>
</tr>
<tr>
<td>$10^{-3}$</td>
<td>9.6</td>
</tr>
<tr>
<td>$10^{-2}$</td>
<td>2.5</td>
</tr>
<tr>
<td>$10^{-1}$</td>
<td>2.5</td>
</tr>
<tr>
<td>$10^{0}$</td>
<td>2.6</td>
</tr>
</tbody>
</table>

Table 5.5 Pulse width dependence of On/Off ratios and memory retention times. $|V_{G2}| = 10V$. 

127
5.4 Simulation results

Figure 5.25 Drain current transient response for the pRAM with $G_1 = 20$ nm and $G_2 = 60$ nm, biased by different pulse widths. Longer pulse widths relate to longer write times, thus less efficient.

**Memory bias dependence**

The magnitude of applied bias (pulse) at the memory contact can determine the occupation of the memory region, but as discussed before, this is primarily by altering the barrier heights at the two gaps, particularly $G_2$. Thus, this factor is expected to have a close relationship to the gap dependence reviewed above. The memory bias dependence is plotted in Fig. 5.26 and the attained changes are

Figure 5.26 Drain current transient response for the pRAM with $G_1 = 20$ nm and $G_2 = 60$ nm, biased by different pulse magnitudes. Increasing the magnitudes achieve long memory retention times for both on and off pulses but not beyond 20 V.
discussed next. Generally, the charged memory time increases for higher magnitude pulses, but the discharged memory time changes very differently. It behaves in a way similar to adjusting $G_2$. A threshold barrier height ($G_2 = 60 \text{ nm}$) was necessary before; and similarly now $V_{G2}$ is required to be at least 10 V for visible discharged memory time. When this bias is increased from 10V (blue) to 20 V (red), the discharged memory time is improved slightly; showing that both memory times are enhanced. However, the pulse of 30 V (green) makes the result completely different without managing to produce any discharged memory time. Moreover, the drain current drops below the equilibrium for 0 V before recovering. The same happens when $G_2$ is increased beyond 80 nm. Hence a correlation can be made as both factors basically affect the barrier height at $G_2$. This occurrence is examined analytically below.

Contour plots of the electron density in the device are shown in Fig. 5.27a for (i) equilibrium before the pulse is applied; (ii) on applying a positive pulse; and (iii) when the pulse ends. The depletion layer in the channel in (i) is reduced during (ii) and at (iii), attempts to recover back to (i). For the $V_{G2} = 30 \text{ V}$ case, during recovery at (iii), the depletion layer seems to be overcompensated. A closer inspection of the conduction channels is performed with zoomed images as a function of time in Fig. 5.27b. A clear depiction of the depletion layer regulating can be seen. At 3 s, when the 30 V pulse is applied, the channel is much more open than with 20 V. At 3.02s, immediately after the pulse is terminated, the recovery to equilibrium is overshot to induce a greater depletion layer than it was before. This would make $I_D$ drop below the intended equilibrium value, Fig. 5.28a (ii). The electron density profiles in the channel at different time intervals confirm the behaviour observed (Fig. 5.28b). The drop in electron density in the channel after the 30 V pulse takes it below the equilibrium (0 V) level, before recovery sets in.

The electron occupation rate in the memory storage is the factor that influences the changes in the channel and is analysed to seek an explanation for the changes. Although the gap size remains constant ($G_2 = 60 \text{ nm}$), the barrier heights are altered considerably for different magnitudes of pulses as depicted in Fig. 5.29. The barrier heights at $G_1$, however, are less influenced (Fig. 5.30). Therefore, at the moment of terminating the 30 V pulse, the effective width at $G_2$ is much greater. The rate of carriers surging into the memory storage region is more substantial as compared to
after the 20 V pulse. This inflow rate is not met by an equal outflow rate due to a much narrower effective gap at G1. A subsequent increased depletion layer is induced in the conduction channel resulting in a drop below equilibrium of $I_D$. It takes ~0.4 s to recover fully back to equilibrium from this situation, Fig. 5.28a (ii).

**Figure 5.27** (a) Electron density contour plots for $V_{G2} = 20$ and 30 V and upon restoring back to equilibrium and (b) electron density contour plots in the channel at different time moments after the positive pulse is removed; for memory biases 20 and 30 V.
5.4 Simulation results

Figure 5.28 (a) Drain current transient response and (b) electron density profiles along the channel at different time moments after the positive pulse is removed; for memory biases 20 and 30 V.

Figure 5.29 Conduction band contour plots showing how they change at G2 for different magnitudes of positive applied bias at the memory gate.
5.4 Simulation results

Figure 5.30 Conduction band energy ($E_c$) probed inside the two gaps as a result of different magnitudes of applied bias at the memory gate.

Exponential fit

A number of parameters have been investigated for the memory device so far, namely: on/off ratio, memory retention time and their dependence on gap sizes, pulse width and memory bias. The parameters that gave a pair of longest memory hold times and the corresponding drain current transient response are shown in Figs. 5.31a and b respectively. The values are multiples of magnitude longer than the refresh rates required of a typical semiconductor DRAM as defined by JEDEC, Foundation for developing Semiconductor Standards, but still short of being nonvolatile. Optimisations of material composition, device design, side-wall coating, and surface treatment, could significantly increase the charge retention time. An improvement in this pRAM though, is that degradation due to charge trapping in the tunnelling oxide is not applicable. Hence the endurance may be higher, where it can withstand more write/erase cycles. The leakage through the gaps by thermionic emission is still possible, reducing memory retention times. High-k dielectrics are an option to reduce this leakage and improve durability. An exponential time decay expression,

$$I = I_o e^{-\frac{t}{\tau}}$$

(5.7)
5.4 Simulation results

fitted to the drain current recovery durations (Figs. 5.31c and d), provides values for their respective decay/growth constants (τ). A higher constant represents a higher rate of thermionic emission. Clearly, the trade-off between ease of charging the memory region and leakage is unavoidable as is generally the case for most memory devices.

Figure 5.31 (a) Summary of the device parameters, (b) its drain current transient, and exponential fits for (c) negative pulse recovery to equilibrium and (d) positive pulse recovery to equilibrium.

5.4.4 Array

Array geometries have proven to be a successful way of enhancing performance of the SSD by stacking ~2000 devices in parallel using an inter-digital structure [49]. In this section, an array of memory storage regions is designed (Fig. 5.32) to observe
5.4 Simulation results

The results are shown in Fig. 5.33. Indeed, the memory time improvements are apparent from the results.

Figure 5.32 Array structures for the pRAM. The numbers represent the number of storage regions, $n$.

Figure 5.33 Current response for array structures. Right: $t_1$ for $n$ number of storage regions.

Though $t_2$ has a slower rate of improvement, the increase in $t_1$ is markedly significant. It may be sufficient as the positive pulse is utilised purely for erasure of the data in flash or for refresh in DRAMs. Hence, the requirement of a long retention time for the charged ‘0’ bit state is more crucial. The table in Fig. 5.33 shows it is possible to derive continual increase in charged memory retention times as the
5.4 Simulation results

number of memory regions increases. Using modern lithography techniques such as nanoimprinting and electron beam lithography, it is quick and easy to produce arrays containing \( n \geq 100 \), where memory times of a few minutes can become possible.

5.4.5 Electrical characteristics

When the negative pulse charges the memory region, the device is in the 0 bit state (written state). Using bias and current values at the drain and the memory contacts, the write and read powers can be calculated for a single memory region as follows:

\[
\text{Power}_{\text{write}} = I_{G2} \times V_{G2} = (-4.0 \times 10^{-10})(-20) = 8.0 \text{ nW}
\]

\[
\text{Power}_{\text{read}} = I_D \times V_D = (7.6 \times 10^{-9})(1.0) = 7.6 \text{ nW}
\]

\[
\text{Energy}_{\text{write}} = \text{Power}_{\text{write}} \times t_{\text{write}} = (8.0 \times 10^{-9})(10^{-3}) = 8 \text{ pJ}
\]

The power and energy values are extremely low, some orders of magnitudes below flash memory devices. The write energies of a typical NAND flash and DRAM are 10 nJ and 2 pJ respectively [234]. Thus, the pRAM established here is potentially valuable to low power portable applications especially in upcoming technologies such as organic, flexible, and printed electronic devices [204].

5.4.6 Oxide-filled trenches

As mentioned in 5.4.1, oxide-filled trenches can bring about increased gate capacitance and potentially reduce operation voltages and even increase memory retention time. This can be practically challenging for very narrow trenches, but simulations can provide predictions for such scenarios. The effect is explored for the device with \( G1 = G2 = 20 \) (longest \( t1 \)). The result is shown in Fig. 5.34. The device
5.5 Discussion and conclusions

with oxide-filled trenches indeed performs better at retaining memory (~50% longer). Furthermore it is at half the bias, confirming the forecasted lower operational power.

![Figure 5.34 Current response for oxide-filled trenches.](image)

5.5 Discussion and conclusions

The boost to memory time has been possible via various adjustments to the electrical inputs such as pulse width and memory bias. However, since simulations provide the facility to test different structures quickly, the preferred method of enhancement is primarily via geometries; so that the best designs can be fabricated, and electrical adjustments can be performed as a second stage tool to improve performances. This has been the objective of this analytical work where both methods were given a good deal of attention so that it may assist experimentalists with valuable information for fabrication and measurements.

The pRAM working principle rests essentially on the gap sizes which modulate the barrier heights in order to regulate drain current. To achieve equivalent memory retention times for both the negative and positive pulse, it was found that three conditions were necessary. Firstly for G1 to be small (~20 nm), secondly for G2 to be larger than G1 (i.e. G1 < G2), and lastly for G2 not to exceed 80 nm (i.e. G2 < 80nm). These conditions provide the opportunity for electron storage time in the memory region to be at a maximum as a result of the effective barrier heights formed by the two gaps. The most ideal combination turned out to be G1 = 20 nm and
5.5 Discussion and conclusions

G2 = 60 nm. Memory achieved by both positive and negative pulses is unique and creates an opportunity for new applications that may require this novelty. At the same time, if nonvolatility is to be achieved, the aspect ratio can be reduced to 1:1 keeping the gap size as small as possible and memory retention times are enhanced, and erasure of information will be quick.

The bias conditions at the memory contact have been studied for pulse width and magnitude. The use of pulse width to improve memory time reaches saturation beyond an optimal duration which was found to be $10^{-3}$ s. The memory bias magnitude applied relates directly to gap size by influencing the barrier heights at the gaps. Therefore, for $G1 = 20$ nm and $G2 = 60$ nm, $V_{G2} = 20$ V produced the most improved times for both $t1$ and $t2$. A higher value, $V_{G2} = 30$ V degraded $t2$ because the barrier height at $G2$ gets lowered more than desired.

Given the simple architecture, it is promising that memory effects can be accomplished in the planar memory structure which can then be densely integrated. An array structure where memory storage regions are placed in parallel was attempted and results show that it is possible to multiply the memory times in relation to the number of regions used. The times achieved shows that the RAM function is readily available, yet the potential to push to nonvolatility is not beyond reach. This is because a small change to barrier height can increase charge retention by order of magnitudes. Continuous design and test, both by simulations and fabrication, can be performed to enhance this possibility. Some possible improvements include using new materials of different carrier concentrations and mobility, optimising device design, employing side-wall coatings, surface treatment and oxide-filled trenches for better insulation and high-k dielectrics for lower leakage.
6

Summary and outlook

6.1 Summary

Semiconductor technology has arrived at a crossroad where device scaling-needs are converging from all facets of the industry. Making significant leaps in the nanoscale regime is paramount to continue the production of high-performance electronics. Reliable and efficient, portable, cheap and simple to construct terahertz frequency emitters, operable at room temperature, are sought for applications in healthcare, security and defence, space, information and communications and others. Device miniaturisation that has continually increased processing speeds in memory applications is reaching fundamental limits. In order to provide solutions to pave the way for future device technologies, this thesis describes two planar nanoelectronic devices. The first makes use of the well-established high electron mobility transistor (HEMT) configuration wherein a two-dimensional electron gas (2DEG) is the conduction layer. An asymmetric conduction channel achieved by etching two L-shaped insulating trenches into the 2DEG layer, produces nonlinearity in the I-V characteristic, thereby acting as a planar diode. Called a self-switching diode (SSD), the device properties and functions have been actively researched over the last decade and well-documented. For example, it has demonstrated rectification, terahertz detection and memory logic-switching. In this work, the SSD’s ability to function as a microwave power generator has been studied. This is an extension from previous predictions using Monte Carlo simulations that showed high frequency emissions from SSDs in the terahertz range.
6.1 Summary

By careful analysis of the electron dynamics in the SSD, the microwave generation has been unambiguously established as Gunn oscillations using ATLAS semiconductor physical modelling. The oscillations are found to originate at a threshold electric field, caused by the electron dipole domain formation and propagation in the conduction channel of the device. The domain dynamics in the active channel gives rise to an oscillating current. The length and width of the channel were found to be critically important in deciding the performance of the Gunn oscillations. Fundamental oscillation frequencies of 0.2 THz were achieved in the shortest channel simulated. Frequency modulation was also possible by adjusting the channel width; showing an inverse relationship. Larger channel widths, however, by virtue of greater electron density, fared better at improving the emission peak magnitude; directly related to the possible emission power. If performances solely depend on geometrical designs, reaching sub-100 nm features can become challenging unless progress is made on the fabrication front. Fortunately, the SSD is a device that is significantly controlled by the surface states that arise due to etching processes. The presence of charge-density can alter the effective length and width of the channel, therefore, providing an alternative route to achieving smaller dimensions. In a device that had a charge density of $1 \times 10^{12}$ cm$^{-2}$, the radiation frequency of higher harmonics could reach 0.7 THz. This is uniquely advantageous since in many conventional devices, efforts are made to eliminate these charges by means of encapsulation and/or passivation. Thus, a combination of carefully selected parameters can enable generation of a broad spectrum reaching 1 THz, which has so far been difficult to achieve in solid-state devices. The selection and tuneability of the frequency and emission magnitude provides a good degree of flexibility in employing such devices in a wide variety of applications.

In a novel approach, a planar random-access memory (pRAM) device is proposed to provide an alternative to conventional vertical CMOS memory structures which are fast approaching their scaling limits. The device is based on the SSD but a charge storage region is introduced next to the conduction channel. The occupation of electrons in the region is controlled by a gate bias thus creating a low current in the channel (state ‘0’) when charged and a high current (state ‘1’) when discharged. Two nanogaps along the perimeter of the memory storage region is also defined to act as adjustable potential barriers to dictate the electron in- and out-flow.
6.2 Outlook

ATLAS simulations facilitate the characterisation of this completely new structure. The simulation is based on a SOI type two-dimensional silicon wafer comparable to current technologies. Initially, I-V and transfer characteristic tests are performed to determine device dimensions. The challenge was to find a channel width that demonstrated a significant gating effect having a high on/off ratio. This width was defined to be 180 nm. Subsequently, the critical nanogaps size dependence was analysed using transient simulations, plotting their memory retention times. The charge retention duration was in the order of $10^{-1}$ s and the on/off ratio was $\sim 10^6$.

The memory retention capability was found to be a function of the potential barriers imposed by the nanogaps. Since the surface-states at the interface boundaries around the nanogaps can be charged and discharged, potential barriers were sensitive to control gate bias. Thus, it was conclusive that encapsulation of charges can be achieved by conscientious engineering of these nanogaps. The optimal pulse width and control gate bias was also determined to be $10^{-3}$ s and 20 V respectively, for the geometry used in the study. Additionally, an array of memory storage regions stacked in parallel was explored. As expected, the memory time scaled linearly with the number of storage regions. This is promising and meets the objective of this work because it would be possible to construct as many storage regions as required with cost-effective single-step lithography, overcoming the scalability issue in vertically-based devices. In terms of energy consumption, the write energy of 8 pJ is in the same order magnitude as DRAMs and three order magnitudes less than flash memory devices. Thus, the pRAM has high value for low-power portable applications, such as organic, flexible, printed, and many other consumer electronics. Given its simple architecture, it is envisaged that design enhancements can develop its memory capacity further, possibly reaching the nonvolatile regime.
6.2 Outlook

Since different frequencies can be achieved by tailoring the dimensions of the SSD, it will be possible to fabricate SSDs in parallel that generate different oscillation frequencies on a single chip making it useful for broadband applications. Therefore, the studies performed here can be extended to array configurations containing SSDs of differing geometrical parameters, which would greatly enrich the information obtained so far. Also, the radiation which occurs in the SSD is along the normal direction of the device surface due to the in-plane oscillating electric field, which would enable coupling of a suitable antenna to the device to assist the radiation. This option can be exercised to amplify and direct the emission power accordingly.

Opportunities to work with different materials for both the THz emitter and the pRAM make prospects a lot brighter than they already are. Gallium Nitride (GaN), often the material of choice for higher frequency emissions in solid-state devices have heat dissipation issues that remain unresolved when used in three-dimensional structures. The employment of GaN to construct SSDs can alleviate the problem and the benefit derived is applicable to many more materials such as CdTe, ZnSe, and GaAsP whose transferred electron effects are being researched for use at higher frequencies [106]. Thus, by the simple task of switching to another material accompanied by SSD array designs, attaining frequencies in the 1-10 THz domain is within reach. The inquest into high-performance CMOS applications has proposed the use of high-k/metal-gate in emerging nanoelectronic devices [235]. These gate electrodes come much closer to achieving the ideal work function required for high performance and low power logic applications. Equipping the pRAM with such gate electrodes can certainly propel the development of the device. Moreover, the use of oxide-filled trenches studied in this work achieved longer memory retention time, so working toward this direction of insulating the trenches with modern high-k dielectric materials should be explored. Another option would be to study the pRAM on a metal oxide layer, which would provide the opportunity to build cheap multilayer high density memory chips in the long run. Finally, it was found that in theory, a rise of 0.5 eV in the barrier height is sufficient to reduce leakage current across it by a factor of $10^8$. This is extremely promising for the development of the pRAM to charter into the nonvolatile regime.
Bibliography

[7] [www.teraview.co.uk](http://www.teraview.co.uk) (20th December 2012).
Bibliography

Bibliography

[45] C. Li, et al., "Design, fabrication and characterization of In0.23Ga0.77As-channel planar Gunn diodes for millimeter wave applications," Solid-State Electronics, 64, 67 (2011).


Bibliography


Bibliography


Bibliography


Bibliography

Appendix A

Gunn Diode Simulation Code

The major components of the SILVACO code used are presented for a 1.2 µm long and 100 nm wide active channel. The statements with hash (#) are remarks which are ignored during simulation. A line continuation operator ‘\’ shows that code continues in the next line.

# SECTION 1: Mesh Input

x.m l=0 spacing=0.05
x.m l=0.05 spacing=0.05
x.m l=0.2 spacing=0.005
x.m l=0.25 spacing=0.005
x.m l=1.55 spacing=0.005
x.m l=1.6 spacing=0.005
x.m l=1.85 spacing=0.05
x.m l=1.9 spacing=0.05

y.m l=0 spacing=0.1
y.m l=0.17 spacing=0.01
y.m l=0.2 spacing=0.05
y.m l=0.205 spacing=0.005
y.m l=0.315 spacing=0.005
y.m l=0.33 spacing=0.02
y.m l=0.35 spacing=0.01
y.m l=0.52 spacing=0.01

# SECTION 2: Structure Specification

region num=1 material=InGaAs x.comp=0.47
region num=2 material=air y.min=0 y.max=0.17 x.min=0.3 x.max=0.5
region num=2 material=air y.min=0.17 y.max=0.21 x.min=0.3 x.max=1.5
region num=2 material=air y.min=0.31 y.max=0.35 x.min=0.3 x.max=1.5
region num=2 material=air y.min=0.35 y.max=0.52 x.min=0.3 x.max=0.5

# Electrodes

electrode num=1 name=cathode x.min=0 x.max=0.05 y.min=0 y.max=0.52
electrode num=2 name=anode x.min=1.85 x.max=1.9 y.min=0 y.max=0.52

# Doping

doping uniform n.type conc = 1e17

# Interface-charge density

interface qf=-4e11
Appendix A: Gunn diode simulation code

# SECTION 3: Models & Material parameters
models material=InGaAs srh fldmob evsatmod=1 \ hcte.el boltzmann temperature=300 print material material=InGaAs mun0=12000 vsat=2.e7 taun0=1.e-8 mobility ecritn=4e3 method newton itlimit=50 carriers=1 output e.field e.velocity e.temp e.mobility

# SECTION 4: DC I-V Simulation
#Initial solution
solve init
solve vanode=0
save outf=sc4_1.2_100_0V.str

#Solve negative bias
solve vanode=-0.1 vstep=-0.1 vfinal=-2 name=anode
save outf=sc4_1.2_100_-2V.str

#Negative to positive bias sweep
log outf=sc4_1.2_100_IV.log
solve vanode=-2 vstep=0.1 vfinal=2 name=anode
save outf=sc4_1.2_100_2V.str

# SECTION 5: Time Domain Simulation
# 2V Transient switching to 2.5V Transient
solve init

#Ramp to 2V
solve vanode=2 ramptime=0.5e-10 tstop=0.5e-10 dt=1e-12

#Solve transient 2V
log outf=sc4_1.2_80_2to3V_osc.log
solve tstop=1.5e-10 dt=5e-13 t.save=5e-13 outfile=sc4_1.2_80_2V_a0 master

#Solve transient 2.5V
solve vanode=2.5 tstop=2.5e-10 dt=5e-13 t.save=5e-13 outfile=sc4_1.2_80_2.5V_a0 master

log off
quit

################################################################END################################################################

N.B. The output statement shown in Section 3 saves additional parameters as required by the user on top of the default ones in the structure file. Additionally, statements such as PROBE were applied for subsequent analysis at specific points/lines in the structure; e.g.:
probe name=efield1 x=0.3 y=0.26 field
probe name=vel1 x=0.3 y=0.26 vel.electron
probe name=efield2 x=1.5 y=0.26 field
probe name=vel2 x=1.5 y=0.26 vel.electron
Appendix B

pRAM Simulation Code

The major components of the SILVACO code used are presented for a 1 µm long and 180 nm wide active channel. The statements with hash (#) are remarks which are ignored during simulation. A line continuation operator ‘\’ shows that code continues in the next line.

# SECTION 1: Mesh Input

mesh
x.m l=0 spacing=0.1
x.m l=0.2 spacing=0.1
x.m l=0.5 spacing=0.05
x.m l=1 spacing=0.01
x.m l=1.2 spacing=0.01
x.m l=1.7 spacing=0.05
x.m l=2 spacing=0.1
x.m l=2.2 spacing=0.1

# y.m l=0.0 spacing=0.1
y.m l=0.2 spacing=0.1
y.m l=0.96 spacing=0.01
y.m l=1.14 spacing=0.01
y.m l=1.8 spacing=0.02
y.m l=1.85 spacing=0.01
y.m l=2.05 spacing=0.01
y.m l=2.1 spacing=0.02
y.m l=3 spacing=0.1

# SECTION 2: Structure Specification

region num=1 material=Silicon
region num=2 material=air y.min=0 y.max=1.02 x.min=0.6 x.max=0.8
region num=2 material=air y.min=1.2 y.max=1.9 x.min=0.6 x.max=0.8
region num=2 material=air y.min=2 y.max=3 x.min=0.6 x.max=0.8
region num=2 material=air y.min=0.82 y.max=1.02 x.min=0.8 x.max=1.4
region num=2 material=air y.min=1.2 y.max=1.4 x.min=0.8 x.max=1.4
region num=2 material=air y.min=2.4 y.max=2.6 x.min=0.8 x.max=1.05
region num=2 material=air y.min=2.4 y.max=2.6 x.min=1.15 x.max=1.4
region num=2 material=air y.min=0 y.max=1.02 x.min=1.4 x.max=1.6
region num=2 material=air y.min=1.2 y.max=3 x.min=1.4 x.max=1.6
Appendix B: pRAM simulation code

# Electrodes

electrode num=1 name=source x.min=0 x.max=0.1 y.min=0 y.max=3
electrode num=2 name=drain x.min=2.1 x.max=2.2 y.min=0 y.max=3
electrode num=3 name=gate x.min=0.8 x.max=1.4 y.min=2.9 y.max=3
electrode num=4 name=anode x.min=0.8 x.max=1.4 y.min=0 y.max=0.1

# Doping

doping uniform n.type conc = 1e17

# Interface-charge density

interface qf=-1e12

# SECTION 3: Models

model Boltzmann conmob srh temperature=300
method newton
output con.band val.band e.mobility

# SECTION 4: Output Characteristics

#Initial solution
solve init

#Vanode solutions: saving temporary files
solve vanode=-10 outf=solve_tmp-10
solve vanode=-5 outf=solve_tmp-5
solve vanode=0  outf=solve_tmp0
solve vanode=5  outf=solve_tmp5
solve vanode=10 outf=solve_tmp10

#load in temporary files and sweep Vd

#Vanode=-10V
load infile=solve_tmp-10
log outf=180nm_output_anode-10.log
solve name=drain vdrain=-20 vfinal=20 vstep=1

#Vanode=-5V
load infile=solve_tmp-5
log outf=180nm_output_anode-5.log
solve name=drain vdrain=-20 vfinal=20 vstep=1

#Vanode=0V
load infile=solve_tmp0
log outf=180nm_output_anode0.log
solve name=drain vdrain=-20 vfinal=20 vstep=1

#Vanode=5V
load infile=solve_tmp5
log outf=180nm_output_anode5.log
solve name=drain vdrain=-20 vfinal=20 vstep=1

#Vanode=10V
load infile=solve_tmp10
log outf=180nm_output_anode10.log
solve name=drain vdrain=-20 vfinal=20 vstep=1
# SECTION 5: Transfer Characteristics

solve init

#Vgate=0V
solve vsource=0
solve vanode=0 vstep=-1 vfinal=-10 name=anode
solve vdrain=1
log outf=180nm_transfer_gate0V.log
solve vanode=-10 vstep=1 vfinal=10 name=anode
log off

#Vgate=-10V
solve vsource=0
solve vanode=10 vstep=-2 vfinal=-10 name=anode
solve vgate=0 vstep=-1 vfinal=-10 name=gate
log outf=180nm_transfer_gate-10V.log
solve vanode=-10 vstep=1 vfinal=10 name=anode
log off

# SECTION 6: Time Domain Simulation

#Initial solution
solve init
solve vanode=0
solve vgate=0 vdrain=1

#Solve transient
log outf= W180_gl_100_g2_100_vg10_pw1e-3.log
solve tstop=1 dt=0.25

#Charging
solve vgate=-10 ramptime=1e-6 tstop=1.001 dt=1e-4
save outf=W180_gl_100_g2_100_180nm_charged_pw1e-3.str
solve vgate=0 tstop=3 dt=0.01

#Discharging
solve vgate=10 ramptime=1e-6 tstop=3.001 dt=1e-4
save outf=W180_gl_100_g2_100_180nm_discharged_pw1e-3.str
solve vgate=0 tstop=5 dt=0.01

log off
quit

###########################################################################END###########################################################################