ULTRA**RAM**™: Design, Modelling, Fabrication and Testing of Ultra-Low-Power III-V Memory Devices and Arrays

Lancaster University

Author:
Dominic Lane

Supervisor:
Prof. Manus Hayne

25th October 2021

A thesis submitted for the degree of:

Doctor of Philosophy

Abstract

In this thesis, a novel memory based on III-V compound semiconductors is studied, both theoretically and experimentally, with the aim of developing a technology with superior performance capabilities to established and emerging rival memories. This technology is known as ULTRA**RAM**™.

The memory concept is based on quantum resonant tunnelling through InAs/AlSb heterostructures, which are engineered to only allow electron tunnelling at precise energy alignment(s) when a bias is applied. The memory device features a floating gate (FG) as the storage medium, where electrons that tunnel through the InAs/AlSb heterostructure are confined in the FG to define the memory logic (0 or 1). The large conduction band offset of the InAs/AlSb heterojunction (2.1 eV) keeps electrons in the FG indefinitely, constituting a non-volatile logic state. Electrons can be removed from the FG via a similar resonant tunnelling process by reversing the voltage polarity. This concept shares similarities with flash memory, however the resonant tunnelling mechanism provides ultra-low-power, low-voltage, high-endurance and high-speed switching capability.

The quantum tunnelling junction is studied in detail using the non-equilibrium Green's function (NEGF) method. Then, Poisson-Schrödinger simulations are used to design a high-contrast readout procedure for the memory using the unusual type-III band-offset of the InAs/GaSb heterojunction. With the theoretical groundwork for the technology laid out, the memory performance is modelled and a high-density ULTRA**RAM**™ memory architecture is proposed for random-access memory applications. Later, NEGF calculations are used for a detailed study of the process tolerances in the tunnelling region required for ULTRA**RAM**™ large-scale wafer manufacture.

Using interfacial misfit array growth techniques, III-V layers (InAs, AlSb and GaSb) for ULTRA**RAM**™ were successfully implemented on both GaAs and Si substrates. Single devices and 2×2 arrays were then fabricated using a top-down processing approach.

The memories demonstrated outstanding memory performance on both substrate materials at 10, 20 and 50 µm gate lengths at room temperature. Non-volatile switching was obtained with $\leq$ 2.5 V pulses, corresponding to a switching energy per unit area that is lower than DRAM and flash by factors of 100 and 1000 respectively. Memory logic was retained for over 24 hours whilst undergoing over 10^6 readout operations. Analysis of the retention data suggests a storage time exceeding 1000 years. Devices showed promising durability results, enduring over 10^7 cycles without degradation, at least two orders of magnitude improvement over flash memory. Switching of the cell's logic was possible at 500 µs pulse durations for a 20 µm gate length, suggesting a sub-ns switching time if scaled to modern-day feature sizes. The proposed half-voltage architecture is shown to operate in principle, where the memory state is preserved during a disturbance test of $> 10^5$ half-cycles. With regard to the device physics, these findings point towards ULTRA**RAM**™ as a universal memory candidate. The path towards future commercial viability relies on process development for aggressive device and array-size scaling and implementation on larger Si wafers.

I hereby declare that, except where specific reference is made to the work of others, the contents of this thesis is my own work and has not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other institute of learning.

Dominic Lane
August 2021

Acknowledgements

The success of ULTRA**RAM**™ as presented in this thesis would not have been possible without the support of my colleagues at Lancaster University.

First, I would like to acknowledge my supervisor, Prof. Manus Hayne for his excellent guidance, patience and contagious optimism throughout the project. In particular, in the many difficult months of process development - we got there in the end! My secondary supervisor, Dr. Andrew Marshall, has also provided crucial support and advice when needed most. I would also like to thank Dr. Peter Hodgson, who has worked tirelessly on improvements to the MBE growth of the memory wafers which was essential to the success of the project.

I would like to thank the cleanroom support staff at Lancaster Quantum Technology Centre: Dr. Chris Somerton, Garry Vernon and Dr. Kunal Lulla for their technical support and guidance with the demanding experimental work in this thesis. I would like to express by sincere gratitude to my ex-colleague, Dr. Ofogh Tizno, who shared her immense knowledge of semiconductor processing and provided training on the equipment needed.

I express my gratitude to Dr. Tom Wilson, Dr. Richard Beanland (University of Warwick), Dr. Adam Craig, Dr. Sergey Kafanov, Dr. Sara Baldock, Dr. Laura Hanks, Dr. Alex Jones and Dr. Alex Robson for their support, assisting with various process and/or characterisation techniques and participating in useful discussions. Further, I would like to thank Dr. Richard Potter (University of Liverpool) who kindly offered to provide initial ALD processing for gate dielectrics after the project had come to halt due to equipment limitations at Lancaster: your help avoided significant disruption to our research.

All of my family, and especially my parents, have shown a great amount of support and patience over the past four years, and to them I am extremely grateful.

Finally, I would like to give special thanks to my wife, Bryanni, who has always encouraged and supported me, and put up with my endless ramblings when sharing our home office during the lengthy coronavirus lockdowns.

List of Publications and Presentations

Articles

- P.D. Hodgson, D. Lane, P.J. Carrington, E. Delli, R. Beanland and M. Hayne 'ULTRARAM: a low-energy, high-endurance, compound-semiconductor memory on silicon,' **Submitted** to Advanced Electronic Materials, 2021

- D. Lane and M. Hayne. 'Simulations of Ultralow-Power Nonvolatile Cells for Random-Access Memory,' *IEEE Transactions on Electron Devices,* 67(2), 474-480, 2020.

- D. Lane, P. D. Hodgson, R. J. Potter, R. Beanland and M. Hayne, 'ULTRARAM: Toward the Development of a III–V Semiconductor, Non-volatile, Random Access Memory,' *IEEE Transactions on Electron Devices,* vol. 68, no. 5, pp. 2271-2274, 2021.

- D. Lane and M. Hayne, 'Simulations of resonant tunnelling through InAs/AlSb heterostructures for ULTRARAM memory,' *Journal of Physics D: Applied Physics* 54, 355104, 2021.

- D. Lane, P.D. Hodgson, R.J. Potter, R. Beanland and M. Hayne, 'Demonstration of a Fast, Low-voltage, III-V Semiconductor, Non-volatile Memory,' *2021 5th IEEE Electron Devices Technology & Manufacturing Conference* (EDTM), Chengdu, 2021.

Patent-pending

- D. Lane and M. Hayne, 'IMPROVEMENTS RELATING TO ELECTRONIC MEMORY DEVICES,' WO/2020/240186 (2020).

Conferences

- Oral presentation at 44th Workshop on Compound Semiconductor Devices and Integrated Circuits held in Europe (WOCSDICE), June 2021 entitled'Non-volatile III-V ULTRARAM memory on Si'.

- Oral Presentation at UK Semiconductors conference, Sheffield, UK (2018) entitled 'Simulations of non-volatile and ultra-low-power resonant tunnelling NVRAM cells'.

- Poster presentation at IEEE Electron Devices Technology and Manufacturing (EDTM) conference, Chengdu, China (2021) entitled 'Demonstration of a Fast, Low-voltage, III-V Semiconductor, Non-volatile Memory' (online participation due to COVID-19).

- Poster presentation at APL Material Challenges for Memory (MCfM) conference 2021 entitled 'III-V non-volatile ULTRARAM™ 2x2 memory arrays' (online).

Contents

Glossary of Abbreviations and Acronyms

CPU	Central processing unit
SRAM	Static random access memory
DRAM	Dynamic random access memory
NAND	A Boolean operator which gives the value zero if and only if all the operands have a value of one, and otherwise has a value of one (equivalent to NOT AND).
IT	Information technology
IoT	Internet of things
SSD	Solid state drive
HDD	Hard disk drive
RAM	Random access memory
AI	Artificial intelligence
MBE	Molecular beam epitaxy
MOS	Metal oxide semiconductor
TTL	Transistor-transistor logic
MOSFET	Metal oxide semiconductor field effect transistor
P	Program
E	Erase
WL	Wordline
BL	Bitline
CMOS	Complementary metal oxide semiconductor
1T1C	one-transistor-capacitor (per bit)
EEPROM	Electrically erasable and programmable read only memory
NOR	A Boolean operator which gives the value one if and only if all operands have a value of zero and otherwise has a value of zero.
FGMOSFET	Floating gate metal oxide semiconductor field effect transistor
FET	Field effect transistor
FG	Floating gate
1T	One transistor (per bit)
FN	Fowler-Nordheim
HEI	Hot electron injection
SILC	Stress induced leakage current
SST	String select transistor
PCM	Phase change memory
CD-RW	Compact disk re-writeable
BJT	Bipolar junction transistor
ReRAM	Resistive random access memory
1T1R	One-transistor-one-resistor (per bit)
1R	One resistor (per bit)
CBRAM	Conductive bridge random access memory
FeRAM	Ferroelectric random access memory
FeFET	Ferroelectric field effect transistor
MRAM	Magnetic random access memory
MTJ	Magnetic tunnel junction

SST-RAM	Spin-torque transfer random access memory
UHV	Ultra-high vacuum
MOCVD	Metal oxide chemical vapour deposition
RHEED	Reflection high-energy electron diffraction
SEM	Scanning electron microscopy
EDX	Energy-dispersive X-ray spectroscopy
AFM	Atomic force microscopy
STM	Scanning tunnelling microscopy
BEXP	Beam-exit cross sectional polishing
TEM	Tunnelling electron microscopy
UV	Ultraviolet
ICP	Inductively coupled plasma
RF	Radio frequency
ALD	Atomic layer deposition
CVD	Chemical vapour deposition
TMA	Trimethylaluminium
PECVD	Plasma enhanced chemical vapour deposition
C-V	Capacitance-voltage
LCR	Inductance capacitance resistance
SMU	Source measure unit
SPICE	Simulation program with integrated circuit emphasis
NEGF	Non-equilibrium Green's functions
RTD	Resonant tunnelling diode
QCL	Quantum cascade laser
HEMT	High electron mobility transistor
TFET	Tunnelling field effect transistor
CB	Conduction band
VB	Valence band
DOS	Density of states
TBRT	Triple barrier resonant tunnelling
QW	Quantum well
RT	Resonant tunnelling
WF	Wavefunction
CG	Control gate
BG	Back gate
S	Source
D	Drain
VCCS	Voltage controlled current source
MSB	Multi-scattering Büttiker
IMF	Interfacial misfit array
DF	Dislocation filter
TLM	Transmission line measurement
TMAH	Tetramethylammonium hydroxide
BOE	Buffered oxide etchant
TDMA-Hf	Tetrakis(dimethylamino)hafnium(IV)
HF	Hydrofluoric
ML	Monolayer (one lattice constant)

Chapter 1

Introduction

The relentless increase in data usage and necessity for improved electronic device performance has placed a significant need for advancements in computer memory technologies to cope with this demand. The memory technology outlined in this thesis, known as ULTRA**RAM**™, achieves the contradictory requirements of a robust memory state with a very low switching energy [1]. This is realised by exploiting quantum-mechanical phenomena achieved using InAs/AlSb and InGaAs/GaSb heterostructures with specific material layer thicknesses.

1.1 The Memory Hierarchy

Modern computers are based on a stored-program concept introduced by John Von Neumann [2]. Programs and data are stored in a separate memory storage unit called memories where information (inputs, outputs, program data and instruction data) are shuttled to and from the central processing unit (CPU) via data buses [3]. In this architecture, the speed at which the memory can be accessed is fundamental to the overall performance of the system. For this reason, different levels of memory are implemented based on their cost, performance and specific architecture, giving rise to a hierarchy of memory technologies (Fig. 1.1).

Static random-access memory (SRAM), dynamic-RAM (DRAM) and NAND-flash are the three technologies favoured for registers and cache, main memory and mass storage respectively (Fig. 1.2). Registers are memory located within the processor which are used to process data and instructions at very high speed. Cache memory acts as a buffer in between the main memory and the CPU registers, temporarily holding copies of information which is frequently accessed by the CPU from the main memory. The main memory is slower than cache but has a larger storage capacity, allowing the working space for the CPU within the Neumann architecture [4]. The final class of memory, at the bottom of the hierarchy, is mass storage. This is where large files and programs are stored and it is generally the slowest memory in the system. SRAM and DRAM are volat-

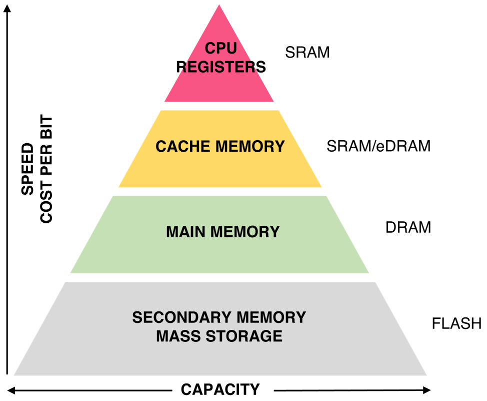


Figure 1.1: The memory hierarchy: the first three memory levels are currently volatile memory (SRAM and DRAM).

ile memories, meaning that their stored data is lost when power is interrupted. Flash, however, is non-volatile. Thus, information can be stored indefinitely. A universal memory seeks to combine all the advantages of each memory class whilst eliminating all of the disadvantages. Such a memory should have the cost, storage-time and scalability of flash memory (*i.e.* non-volatile and large capacity), and the switching speed and energy requirements of DRAM/SRAM. ULTRA**RAM**™ possesses many of these qualities, with the added benefit of having an extraordinarily low switching energy and a non-destructive read.

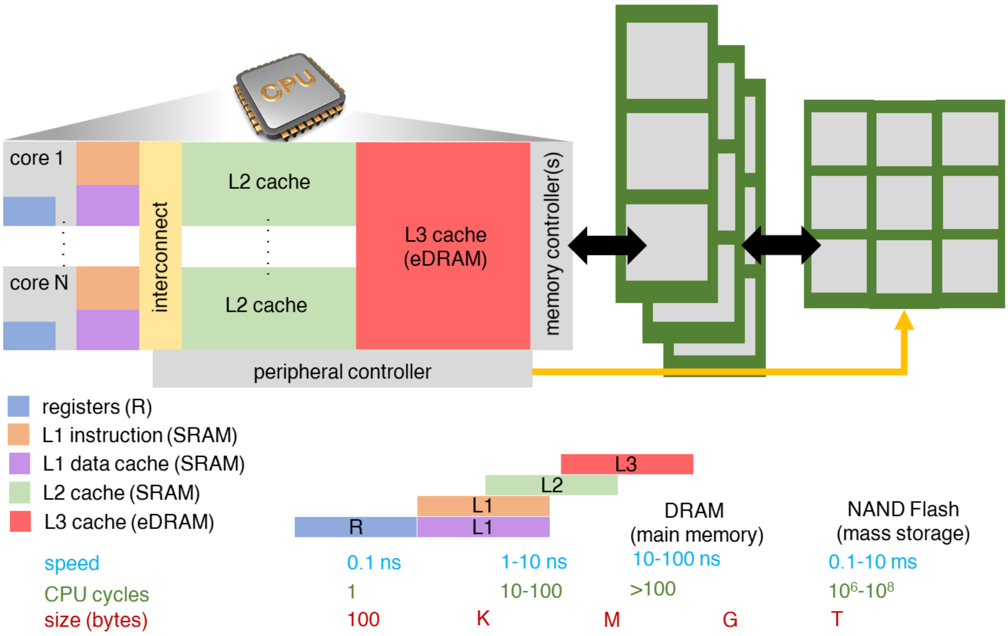


Figure 1.2: Overview of commercial memory technologies and their position in the memory hierarchy. Adapted from [1].

1.2 The Memory Wall

Dynamic random-access memory (DRAM) represents 99% of RAM used in electronics today [5]. As CPU performance continues to advance according to Moore's law [6] and Dennard scaling[1] [7], the memory which provides the platform for these super-powered processors is struggling to keep up. This is due to some fundamental limitations of DRAM, which are discussed in detail in subsection 2.1.2 of the following chapter. Known as the memory wall, or the processor-memory performance gap, it is an overall drawback in computer performance and is now the primary obstacle to improved computer system performance [8].

With many clock cycles per memory access, increasing processor clock speed is no longer a gateway to increased performance [1]. The industry has focused on developing different levels of caching, improving bus controllers and implementing prefetching techniques to try to mitigate the processor-memory disparity. However, it is clear that the gap is continuing to grow and a significant amount of resources are being used to search for a brand-new memory technology to replace DRAM [9, 10]. Based on the evidence presented in this thesis, ULTRA**RAM**™ could provide significant improvements in performance, capacity and energy efficiency compared to DRAM, thus broadening the memory bottleneck and maximising the strides made in other areas of computing.

1.3 Power Consumption and the Internet of Things

The recent explosion of internet traffic and production of new electronic devices has led to a dramatic rise in energy consumption associated with information technology (IT). Models suggest that IT will account for 21% of global electricity demand by 2030, with a significant share of energy being used for networks and colossal data centres [11]. Currently, if the global IT industry were a country, only China and the United States would contribute more to climate change [12]. This figure is poised to increase sooner rather than later, particularly if computationally intensive cryptocurrencies such as bitcoin continue to grow [13]. Most of the largest data centres are in hot or temperate climates, consuming vast amounts of energy to keep them from overheating. Storing, moving, processing, and analysing data all require energy. Typically these data centres store the information in solid state drives (SSDs) or magnetic hard disk drives (HDDs) and are processed using DRAM as the working memory [14]. DRAM consumes more than 25% of datacenter energy and contributes to the necessary cooling overhead, which accounts for another 15% of energy usage [14]. Introduction of ULTRA**RAM**™ into this sector, even if only as a RAM, would provide tremendous efficiency gains which would greatly reduce the environmental impact of Netflix® binge-watching [15].

[1] Often folded in to Moore's law, but discovered separately by Robert H. Dennard in 1974, is the observation that processor performance per watt grows at this same rate as Moore's law, doubling about every two years.

It is predicted that we are at the beginning of a ‘Second Machine Age’ [16] or a ‘Fourth Industrial Revolution’ [17]. This refers to the introduction of disruptive technologies and trends such as the Internet of Things (IoT), robotics, and artificial intelligence (AI) changing the way we live and work. IBM’s Watson impressively defeated reigning Jeopardy! champions in 2011, demonstrating that AI computers were better at answering questions posed in natural language than humans. However, Watson consumes 80 kW of power, whereas a human brain typically uses a few tens of Watts [18]. In order for AI machines to become commonplace in our daily lives, there would have to be a large shift in system and training efficiency [20]. Considering Watson’s 16 TB of RAM [19], an ultra-low-power memory would provide significant progress towards this goal.

The IoT refers to a system of interrelated computing devices, mechanical and digital machines, objects, animals or people with the ability to transfer data over a network without requiring human-to-human or human-to-computer interaction [21]. The implementation of such a system has some fundamental and practical challenges which exist at the level of the individual sensor and associated hardware embedded in the ‘Things’. In particular, many of these sensors may be electrically isolated (autonomous), either having their own power source (battery) and/or be required to harvest or scavenge energy from their environment, for example from solar, wind, thermal or vibrational sources. Harvesting provides very little energy, so such autonomous sensors are required to have the lowest possible power consumption. They may even need to remain in a quiescent state slowly collecting or storing data for extended periods, before bursting into life when sufficient energy is available to connect to the outside world [21]. A key component in achieving this is the devices’ memory. Such a memory should work at low voltages, use as little energy as possible when being programmed, and be capable of robustly storing the data almost indefinitely with no power.

1.4 Synopsis

This thesis begins with an overview of competing memory technologies and their device physics, both in commercial production and those still under development (Chapter 2), providing a reference point from which we can compare the properties of ULTRA**RAM**™. Chapter 3 provides details of the research methods used to simulate and develop the memory technology and fabricate and characterise memory devices and arrays. The memory concept is introduced in Chapter 4, where the fundamental principles and background theory are first presented before a detailed discussion of device operation. Sections 4.5 and 4.6 present a novel method of simulating memory performance, the results are which are discussed in Section 4.7 and are followed by a complementary high-density RAM architecture realised from the results (4.8).

The next two chapters are the experimental work on fabricated devices. Chapter 5 describes the realisation of the memory designs on GaAs and Si substrates using molecular beam epitaxy (MBE) to grow the III-V layers followed by a top-down fabrication approach, and the electrical results are detailed and discussed in Chapter 6. We then return to more simulation results in Chapter 7, where we use a similar technique to Chapter 4 but with modified layer structures to assess the necessary growth tolerances for ULTRA**RAM**™ memory function. In the final chapter we conclude and discuss the future development of the ULTRA**RAM**™ emerging memory concept.

Chapter 2

Memory Technologies

A memory refers to a device which preserves information for retrieval [22], usually in a machine language. The earliest example of such a device is the punch card. Developed by Basile Bouchon in 1725, punched holes represented a ‘sequence of instructions’ for pieces of equipment, such as textile looms and player pianos [23]. Over 200 years later, Claude Shannon, of Bell Labs, realised the similarities between Boolean algebra and electrical circuits. Consequently, the field of information theory; storing and processing information as binary code which could be electrically implemented, gave birth to the digital revolution [24, 25]. From this point onwards, memory devices have largely referred to electrical and magnetic technologies capable of storing binary information as sequences of 0’s and 1’s.

This chapter details the memory technologies currently in production and those still in their infancy. Understanding the operation and architecture of the technologies and recognising their advantages and limitations is of vital importance. Firstly, the chapter provides us with the background necessary to compare ULTRA**RAM**™. Secondly, because certain aspects of current technologies have been appropriated in developing the technology, the chapter is a necessary prerequisite for the subsequent chapters of this work.

2.1 Volatile Memories

Volatile memory refers to memory technologies that only maintain their data while the device is powered. Today’s most popular RAM products are SRAM and DRAM based, both of which are volatile memories.

2.1.1 SRAM

In the early days of the Micrologic family development, Bob Norman suggested that multiple semiconductor flip-flops could be used to build a memory array but he ‘decided it was so economically ridiculous, it didn’t make any sense to file a patent on it’ [26]. Just a few years later, in 1964, John Schmidt designed a 64-bit MOS p-channel static random access memory (SRAM) [26]. Soon after, in 1969, Intel released the first SRAM chip, the 64-bit Intel 3101 [27] followed closely by the Intel 1101, a 256-bit version and the first metal-oxide-semiconductor (MOS) product (Fig. 2.1) [28]. They used Schottky TTL (transistor-transistor logic). Although six transistors were required to store each bit (1538 transistors for the Intel 1101 [29]), the semiconductor integrated circuit (IC) platform allowed for monolithic processing and device scaling, and therefore receives performance and cost upgrades as specified by Dennard scaling [7] and Moore’s law [6] respectively. Consequently, the performance and cost per bit of SRAM improved significantly faster than its competitors. By the early ’70’s, SRAM had almost completely eradicated magnetic-core memory, a technology that had dominated the RAM market for almost 20 years [30].

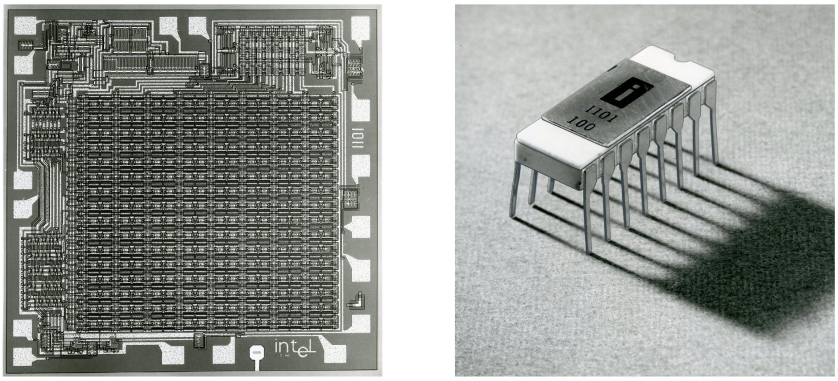


Figure 2.1: The Intel 1101 was the first high-volume MOS memory. Left: Die shot. Right: Packaged product. Credit: Intel Corporation.

Over half a century later, the structure of the individual SRAM bit is largely unchanged from the original Intel 1101. Fig. 2.2 **(a)** shows the TTL SRAM cell consisting of six metal-oxide-semiconductor field-effect transistors (MOSFETs) [31]. The central four transistors, labelled MOS-1 to MOS-4, compose a bi-stable flip-flop circuit which stores the memory state. To demonstrate the operational principles of this circuit, the equivalent logical schematic is provided in Fig. 2.2 **(b)**, which is two connected inverters. The output potential of each inverter is fed as input into the other. This feedback loop stabilizes the inverters to their respective state. For instance, if we work clockwise around the loop, an input of $Q = 1$ inverts to $\overline{Q} = 0$ and back to $Q = 1$ and the loop is formed. MOS-5 and MOS-6 are select transistors with their gate terminals connected to the word-line (WL) and their source terminals connected to adjacent bit-lines (BLs), labelled BL and $\overline{BL}$. By applying a sufficient voltage across the gates of the transistors, the channel of transistors will open [32][1]. This allows us to access the flip-flop circuit by applying a signals along BL and $\overline{BL}$ to read, program (P) or erase (E) the data within. When there is no voltage on the WL, the select transistors provide electrical isolation from the rest of the circuit, such that the inverters sustain the logic state provided there is power supplied to them [31].

In order to program (*i.e.* set to logical state 0) or erase (*i.e.* set the logical state to 1) the bit, a specific signal will be used to disrupt the stability of the feedback loop such that it restabilises with the desired logic [31]. To demonstrate this principle, consider that we put $Q = 1$ and $\overline{Q} = 0$ on BL and $\overline{BL}$ respectively, at precisely the same moment. As this is consistent with the inverter loop, the stability of the state prior to this signal is overwritten and the loop remains in this state.

In order to read the data, the WL is addressed once more to 'open' transistors MOS-5 and MOS-6, giving us access to the logic circuit. The logical states Q and $\overline{Q}$ produce a voltage difference across MOS-6, but not MOS-5, holding V_{dd} on both sides. This, in turn, results in current flow and $C_{\overline{BL}}$ discharge, producing a measurable voltage difference across the two bit-lines with a polarity corresponding to the logical state stored inside [31].

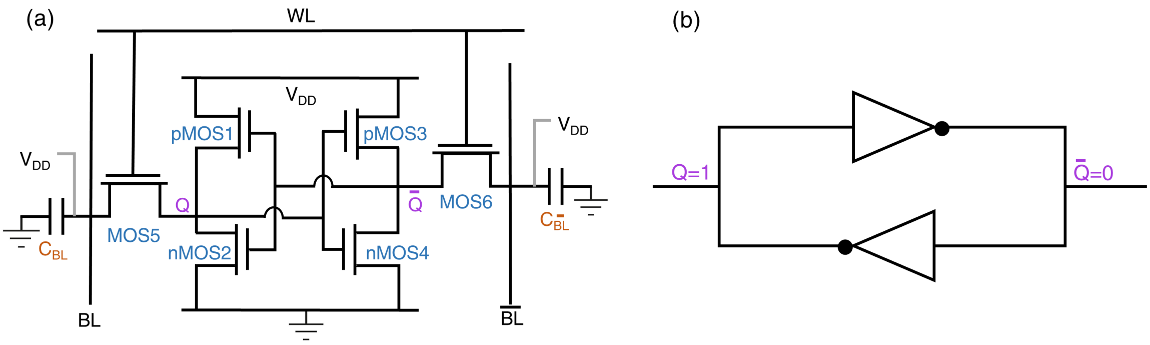


Figure 2.2: SRAM circuit diagrams: **(a)** full 6-transistor architecture and **(b)** simplified inverter schematic.

When the 256-bit Intel 1101 was released, the processing technology of the day was of 12 µm feature size. The size of the transistors were the fundamental limitation in both bit density and speed [6, 7], corresponding to a 850 ns access time [29]. The processing technology of complementary metal-oxide-semiconductor (CMOS) fabrication today is minuscule by comparison, with TSMC and Samsung electronics manufacturing at the 5 nm node, corresponding to an SRAM cell size of 0.017 μm^2 [33, 34]. The relentless cramming of components onto silicon chips meant that, by 1995, SRAM speed was already around 6 ns [35]. SRAM access speeds are now $\sim$ 1 ns, keeping up with the clock speed of the fastest CPUs [1, 36].

In the ever-shrinking world of CMOS fabrication, it can be difficult to keep up with advancements in bit densities. To simplify this, the unit F^2 is used to describe cell area, where 'F' is the

[1] The term 'open' refers to activating the channel (*i.e.* switching the channel into a conducting 'ON' state) of an enhancement-mode MOSFET. This occurs when the applied gate bias overcomes the potential barriers formed by the p-n junctions of the depletion regions, allowing carriers to flow through the channel.

feature size of the process [36]. Therefore today's SRAM technology will require the same number of F^2 in tomorrow's technology. Later, this will allow us to compare the bit density of ULTRA**RAM**™ with current technologies, despite our current inability to fabricate devices at industry-leading node sizes. Although the MOSFETs in current SRAM technology are miniscule, six of them (6T) are required along with two bit-lines (so an extra interconnect), corresponding to a 100 F^2 minimum cell area [36]. For this reason alone, SRAM has been the reserve of high-speed, low-density applications such as CPU registers and caching, as it is the only current memory technology that can be accessed at the tempo of the CPU [36]. The larger capacity required for general RAM in today's computers is almost entirely the role of dynamic-RAM [5], another volatile memory described in the following section.

2.1.2 DRAM

The single transistor dynamic-RAM (DRAM) was invented in 1967 by Dr. Robert Dennard of IBM [37]. The single transistor DRAM consists of a MOSFET connected to a capacitor, as pictured in Figure 2.3. When the capacitor is charged, the memory cell is in state '1'. Conversely, when the capacitor is discharged, the cell is in state '0'. The reading and writing of the cell is controlled through the transistor, acting as a selection device whereby applying a word-line voltage opens the channel of the cell. Consequently, the bitline can be used to flow current into the capacitor by passing electrons through the (now conducting) MOSFET channel. Such a program/erase (P/E) architecture is extremely compact; only a single world-line and bit-line is required to select column and row on the memory cells. This allows program and erasing of devices in any order (*i.e.* at random): a format suitable for random access memory [38, 39].

The readout mechanism for the single-transistor DRAM brings about some difficulties. To read the data stored in the device, we select the relevant wordline to open the MOSFET channel and read the current on the bitline as a result of the capacitor discharging. If there is a current spike, the state was '1', otherwise, it was in state '0'. Naturally, discharging the capacitor means that the data is no longer stored in the cell and has to be re-written. This is known as a destructive-read [38].

As the memory storage is implemented by a charged capacitor, the memory is unalterably volatile (*i.e.* the capacitors discharge as soon as power is interrupted) [38]. Moreover, the capacitors also discharge gradually over time (every 60 ms or so [40]) due to leakage. The result is that constant reading and refreshing (*i.e.* rewriting) of the data is necessary to retain the memory state. In order to deal with the added complexity of this destructive readout, peripheral circuitry must be added to sense which cells need re-programming and which ones do not (*i.e.* if read-out is 1 or 0 respectively). The constant refreshing of data significantly increases power consumption. The power consumption of the refresh is especially problematic in battery-powered devices such as mobile phones, where the ceaseless refreshment of memory cells significantly reduces battery

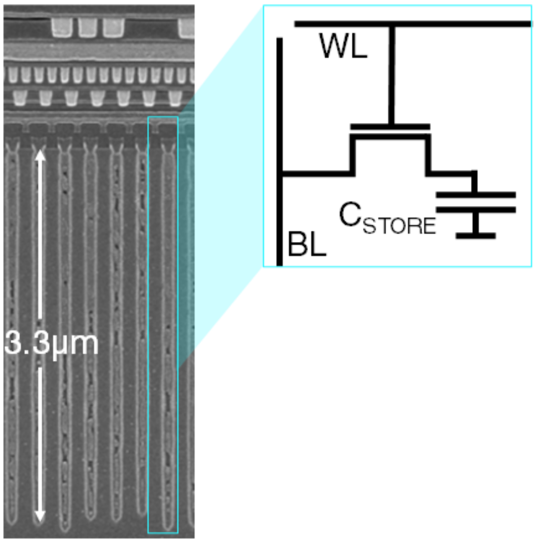


Figure 2.3: Left: memory array of deep trench capacitors[1]. Cross section of the DRAM cells shows the high-aspect ratio capacitors with the select transistors. Right: circuit schematic for the single-MOSFET DRAM cell.

life [41].

The success of DRAM is largely due to its performance (*i.e.* read and P/E speed), scalability (cost and capacity) and endurance. The read and P/E speed of DRAM is typically < 10 ns [36]. Although this is an order of magnitude slower than SRAM, it is significantly faster than other non-volatile and high-capacity alternatives such as flash. It is important to note that the speed limitation is due to the capacitative nature of the memory storage. The speed-limit of DRAM is essentially a result of the *RC* time constant [45], *i.e.* access time, τ is

$$\tau = RC \; , \tag{2.1}$$

where *R* is the circuit resistance and *C* is the capacitance. However, the capacitance of individual cells must be sufficient to exceed the circuit parasitic capacitance in order to distinguish between some parasitic discharge and a memory state upon readout [42], and must be kept large enough to manage refresh frequency. As a consequence, DRAM cells cannot simply have their capacitance reduced to improve performance. Moreover, bit-line and word-line length (*i.e.* number of cells on a given line) are restricted by parasitic capacitances, as when stringing 100's of cells together this problem quickly becomes significant. In fact, DRAM word-lines and bit-lines today are a few thousand cells long (*i.e.* a few MBit array), and the high capacity DRAMs used in electronics consist of many relatively small arrays on the same die [43].

Although capacitance is the limiting factor for DRAM's speed, a high-capacitance memory

[1] Copyright: Chipworks under fair-dealing.

cell is necessary to reduce power consumption[2] [44], and allow for larger arrays of memory cells. The MOSFET part of the one-transistor-one-capacitor (1T1C) cell has been scaled conveniently with the advancements of Moore's law [6], however scaling the capacitor whilst maintaining sufficient capacitance has been an entirely different challenge. This is mainly a consequence of capacitance dependence on area, *i.e.* for a a parallel plate capacitor

$$C = \frac{k\epsilon_0 A}{d} \, , \tag{2.2}$$

where A is cross-sectional area, k is the relative permittivity for the capacitor material, ϵ_0 is the vacuum permittivity and d is the distance between the plates [45]. As the components get smaller, A decreases and the capacitance suffers as a result. To compensate for this problem, DRAM manufacturers implemented downward trench capacitors with extremely high aspect ratio, allowing for the capacitative area to act vertically, rather than occupying valuable space on the silicon surface [46]. An example of this is shown in Fig. 2.3 (left). Another way to increase capacitance is to consider parameters k and d. Nowadays, atomic layer deposition (ALD) is used in commercial DRAM manufacturing. This method, described in detail in subsection 3.3.7 of the following chapter, allows for deposition of high-k dielectric on high-aspect ratio trenches with layer thickness precision on the atomic scale [47, 48]. These advancements have allowed DRAM to be scaled to 14 nm technology [49] with a cell area of 6 F^2[36]. Consequently, DRAM is the fastest memory capable of having a large capacity and relatively-low cost. It therefore remains the primary choice for working memory in electronics, despite its glaring flaws.

2.2 Non-volatile Memories

2.2.1 Flash

Flash memory is a type of electrically-erasable programmable read-only memory (EEPROM)[3]. The name 'flash' refers to the ability to erase all the cells in a block in one operation [50]. The individual devices that constitute the memory array are floating-gate MOSFETS (FGMOSFETs), the same as EEPROM-memory which allowed for byte-level erasing only [51]. Dr. Fujio Masuoka invented flash memory when he worked for Toshiba in the 1980s. Masuoka's colleague, Shoji Ariizumi, reportedly coined the term flash because the process of erasing all the data at once reminded him of the flash of a camera [52]. The flash memory used today can be separated into two categories, NOR-flash and NAND-flash. These memories use FGMOSFETs to store the memory state but implement different array architectures for different applications. In general, NOR-Flash and NAND-flash are used for embedded and high-density storage applications respectively [52].

[2] A higher capacitance cell results in more electrons defining the memory state, therefore the capacitor discharge through leakage will take significantly longer and refresh time is reduced.

[3] This is not actually a read-only memory as the name suggests, as it is electrically erasable and programmable.

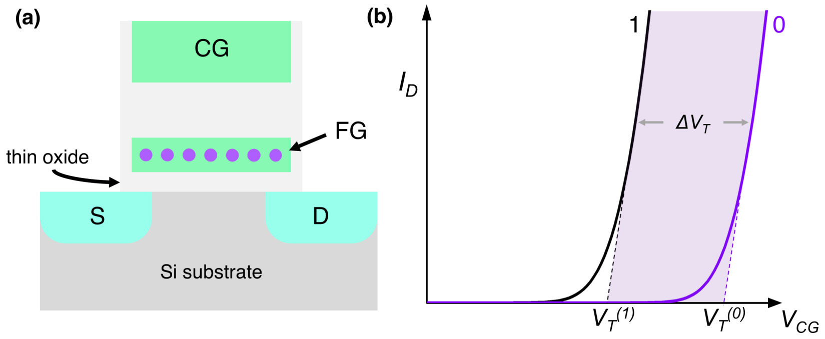


Figure 2.4: **(a)** Schematic of the FGMOSFET used to store memory in Flash arrays. It is a charge-based memory in which electrons are stored on the FG (pictured as purple dots). **(b)** Drain current measurements during a CG voltage sweep for the memory in states 1 and 0.

The FGMOSFET is based on a silicon-MOS structure as depicted in Figure 2.4**(a)**. A floating gate (FG) is formed by a second oxide layer such that the FG is isolated from both the MOS-channel and the outer gate, referred to as the control gate (CG). The memory state is defined by the charge stored in the FG, *i.e.* the absence or presence of electrons in the FG define memory states 1 and 0 respectively [53]. The objective of the P and E cycles is to move electrons from the MOS-channel in to or out of the FG, and for those electrons to remain there indefinitely until the next cycle, even with no power in the system. This is achieved using Fowler-Nordheim tunnelling or hot-electron injection (HEI). A detailed description of these processes is given later in this section.

The read mechanism for the individual FGMOSFET is straightforward. The channel of the FGMOSFET typically forms an enhancement mode FET (using doped junctions), resulting in a threshold voltage (V_T). This is the control gate (CG) voltage at which the channel switches from a high-resistance to low-resistance state, leading to the drain current (I_D)- control gate voltage (V_{CG}) characteristic shown in Fig. 2.4**(b)**, black colour line [53]. The negative charges (electrons, pictured in Fig. 2.4**(a)**) gathered in the FG layer create a negative potential, which screens the applied gate potential when performing a $I_D - V_{CG}$ sweep. Consequently, the threshold voltage of the channel is shifted by the programmed state by

$$\Delta V_T = \frac{Q_{FG}}{C_{FG}} , \tag{2.3}$$

where ΔV_T is the threshold voltage shift, Q_{FG} is the charge stored in the FG and C_{FG} is the capacitance between the CG and FG [54]. This shifting creates what is commonly-known as the threshold voltage window, *i.e.* the gap between the thresholds for the FGMOSFET in the programmed and erased states. If we want to determine the memory state of the FG, we apply a CG voltage within the threshold voltage window. As depicted on the graph in Fig. 2.4, if the FGMOSFET is in the erased state (1), we are able to detect a significant I_D current (*i.e.* ON transistor state). If the FGMOSFET is in the programmed state (0), the charge screening does not allow the MOS-channel to witness a sufficient CG voltage to cross the threshold (*i.e.* OFF transistor state). Consequently, the measured I_D for the logic 0 state is much smaller than for logic state 1. The design of this readout mechanism is advantageous as it is non-destructive and allows for highly scalable devices with just one-transistor per bit (1T) [55].

Fowler-Nordheim tunnelling

Fowler-Nordheim (FN) tunnelling refers to the emission of charge via quantum-mechanical tunnelling as induced by an electrostatic field. This can take place from solid or liquid surfaces, into vacuum, air, a fluid, or any non-conducting or weakly conducting dielectric. For the context of this chapter we will focus on bulk crystalline solids, specifically the silicon-silicon dioxide-poly structure typically used in FGMOSFETs for flash memory[4] [60]. Figure 2.5 depicts the conduction band-structure of a typical device under **(a)** zero electric field and **(b)** high electric field, applied through voltages on the device contacts. At zero field, the SiO_2 barrier provides a robust square barrier sufficient to prevent electrons moving through it. However, under high field, the shift in the conduction band on the dielectric produces a triangular shaped barrier. Under these conditions, tunnelling probability is greatly increased as the barrier thickness near tip of the triangle (Fig. 2.5**(b)**) is significantly thinner by virtue of the barrier shape [61]. The current density (J_{FN}) resulting from the increased tunnelling probability is of the form

$$J_{FN} = \alpha E^2 e^{-\frac{\beta}{E}} \ , \tag{2.4}$$

where α and β are FN parameters dependent on many variables and, as such, are usually determined experimentally [62, 63]. This equation is derived from the Wentzel-Kramers-Brillouin (WKB) approximation, assuming a perfectly triangular barrier.

Typically, FGMOS-devices in flash memory require an input voltage $> 10\ V$[5] to transfer charge across the barrier to program/erase memory states [36]. In designing a memory, it is important to consider that the limitations on the thickness of the tunnelling oxide for an increased J_{FN} and low-voltage operation, as the barrier must provide sufficient resistance to direct tunnelling in order to retain the memory state. Nevertheless, the tunnelling barriers of recent NAND-flash structures are just 6 nm thick [56].

[4] Although this is the typical structure seen in textbooks, recent flash memories use alternative high-k dielectric materials deposited via ALD with metals for both improved performance and the ability to scale to TSMC's 28 nm process or smaller [56, 57, 58, 59].

[5] 18 V in the example of Fig. 2.8.

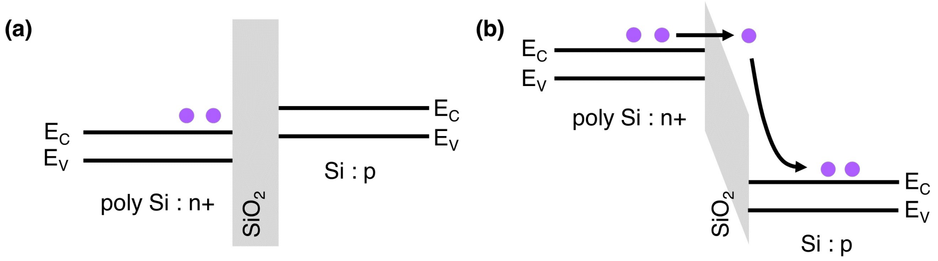


Figure 2.5: Band-structure profile for a poly-SiO_2-Si structure under **(a)** zero electric field and **(a)** high electric field. Electrons are represented by purple dots.

Hot electron injection

Hot[6] electron injection (HEI) is an alternative mechanism which also allows electrons to move over the SiO_2 barrier from the channel-side of the device. HEI occurs when a carrier possesses a kinetic energy (E_K) sufficient to overcome the potential barrier provided by the oxide layer, E_{barrier} [64] *i.e.* when

$$E_K > E_{\text{barrier}} \;, \tag{2.5}$$

whereby

$$E_K = eE_{\perp}\ell \;, \tag{2.6}$$

in which e is electron charge, $E_{\perp}$ is the electric field applied to the perpendicular to the direction of injected carriers (*i.e.* underneath the barrier) and ℓ is the distance between carrier collisions in the channel [64]. HEI is primarily used in NOR-flash as it lends itself to this arrangement for single-bit access, the details of which will be discussed in due course. Due to the dependence on $E_{\perp}$, the required voltage to fulfil Eq. 2.6 is reduced as the length of the underlying channel decreases. In the context of the FG-devices used in flash-memories, the required voltage to program the memory state is reduced as devices scale down [65]. However, for regular MOSFETs, this phenomena is a major factor in gate leakage problems for nanoscale devices [66]. It is important to note that this method can only be used to program the memory state (*i.e.* add electrons to the FG), as it requires electrons to travel perpendicularly to the barrier at high energy. To erase the memory state (*i.e.* remove electrons from the FG) we must revert to FN tunnelling as electrons are localised in the FG as the storage media.

Failure mechanisms

Flash-memory has an endurance of around 10^5 program and erase cycles [36], after which the devices are unable to reliably retain data in their FGs. Consequently, devices can wear out with

[6] The term 'hot' refers to the effective temperature used to model carrier density, not to the overall temperature of the device.

heavy usage. In order to mitigate this problem, chip-producers go to great lengths to evenly distribute writing on all blocks of a SSD so they wear evenly; a process known as wear-levelling [67]. Manufacturers also use ‘bad-block-management’, whereby the memory core contains blocks which are not available to the user (*i.e.* not part of normal storage capacity), but instead are activated when the drive detects a worn out memory block [68]. Despite these efforts, current flash-based storage drives (SSDs) offer a lifespan of 10 years or less [69].

The primary failure mechanisms for the FG memory devices are voltage accelerated failures within the oxide tunnelling barrier. The frequent application of large voltages across the thin oxide barrier creates stress within the structure which adds to the natural traps caused by imperfections in the oxide [70]. This increase in traps allows electrons to flow to or from the FG by trap-assisted tunnelling, known as stress-induced leakage currents (SILC). HEI injection also causes oxide degradation, especially in the vicinity of the drain terminal. This leads to electrons being trapped at the Si/tunnelling-oxide interface and within the oxide layer [71]. Eventually the electrons can be released from the traps causing charge variation and subsequently effect the threshold voltage of the device. FN tunnelling generates negative traps across the entire insulating layer, as well as degradation due to hot-hole generation and trap-to-trap hopping. The oxide breakdown mechanisms which cause memory failure are numerous [72], but all are a direct result of the unavoidably large electric fields required to operate the program/erase cycles.

An alternative flash cell construction is known as charge trap flash (CTF), where the program and erase mechanism is similar, however the conducting FG layer is substituted for an insulating one. This improves the cell’s resilience to SILC, as a short circuit created by voltage-induced defect between the charge trapping layer and the channel will eliminate the stored electrons only in direct contact with the defect, leaving the other electrons in place to continue to control the threshold voltage of the device. The inclusion of the FG insulator allows thinner tunnelling barriers to be used which improves performance and switching energy. Moreover, this construction is favourable for 3D stacking of flash arrays, and is the device of choice of the majority of flash technologies at present[7] [73].

Performance

The speed of the quantum-mechanical tunnelling process is of the sub-ps scale [74]. Considering this, one may predict that flash memory will perform extremely quickly. Disappointingly, this is not the case. The limitation of flash memory performance is a result of the *RC* time constant (Equation 2.1), as discussed in the previous section. With typical device capacitances of around 50 aF [75] at the 20 nm feature size, compared with DRAM cell capacitances of 15 fF for similar scale, this still does not explain the slower-than-DRAM performance of flash memory.

Unlike DRAM, Flash-memory requires high-voltages (>10 V) for P/E cycles. These voltages

[7] Flash devices in the comparison table are CTF cells.

are not readily available on chip. Thus, charge pumps are an indispensable component in increasing the voltage to the desired level [76]. However, the charge pump relies on capacitors for energetic charge storage to raise voltage. Consequently, the speed of program and erasing Flash memory is, in reality, dominated by the *RC* time constant for the capacitors in the charge-pump circuitry. The result is that the Flash-memory has a P/E speed of around 10 µs [75], three orders of magnitude slower than DRAM [36].

NAND-flash

NAND-flash is composed of strings of FGMOSFETs connected in series to form a bit-line (BL), which is controlled by select transistors (MOSFETs) at either end. These are arranged in blocks (Fig. 2.6) which typically consist of up to 64 pages. A page refers to all the bits in the word-line, as pictured in the pink box of Fig. 2.6, and is the smallest granularity of data that can be addressed by the external controller. Page length can be as long as 2,000 bytes (16k bits), corresponding to a block-size of 135 kB for single-level cell flash memory [78].

In the NAND-flash architecture, program cycles are carried out page-wise, whereas erase cycles are carried out block-wise (both via FN tunnelling). To erase the block, a large negative ($>$ -10 V) voltage is placed upon each word-line (WL)[8]. Accordingly, all FGs in the block are simultaneously emptied via FN tunnelling, leaving a block of ‘1’ state memory cells. To program the memory cells page-wise, the string-select transistor (SST) is turned on while the ground-select transistor (GST) is switched off by the control unit. The targeted cells within the page are then programmed by applying a high negative voltage to the word-line, and biasing the bit-lines corresponding to ‘0’ to ground by selectively opening the channels of both the GSTs and other memory cells in the string. Only the cells with a connection to ground will be programmed, as this is required to affix the voltage across the oxide to write the state via FN tunnelling [79]. The result is that NAND-flash can be programmed (*i.e.* setting 0’s) in entire pages, allowing for fast mass memory storage (*i.e.* high bandwidth) [80].

To read the memory state, we apply a gate voltage (on the target WL) with a value within the threshold voltage window (Fig. 2.4) such that, if the channel is in state ‘1’, the voltage is sufficient to significantly enhance the conductivity of the channel. Otherwise, its memory state is ‘0’. Simultaneously, we apply a voltage to all other memory cells in the same string (on WLs), such that they conduct regardless of their memory state, *i.e.* $V_{CG} > V_{TH}^{(0)}$ (Fig. 2.4). A voltage is also applied to the select transistors to provide full access to the string. Now, measuring the BL current; if the cell is in state ‘1’, the threshold window voltage will allow a conducting path from BL to ground such that a current is measured. If the same cell is in state ‘0’, the channel will provide a high resistance, blocking the flow of current through the string. Thus, we can determine the memory state of an individual cell by measuring the current through an entire string (BL) [79].

[8] Alternatively, a block can be erased using a positive potential on the device base, resulting in the same electric field across the tunnelling barrier [79].

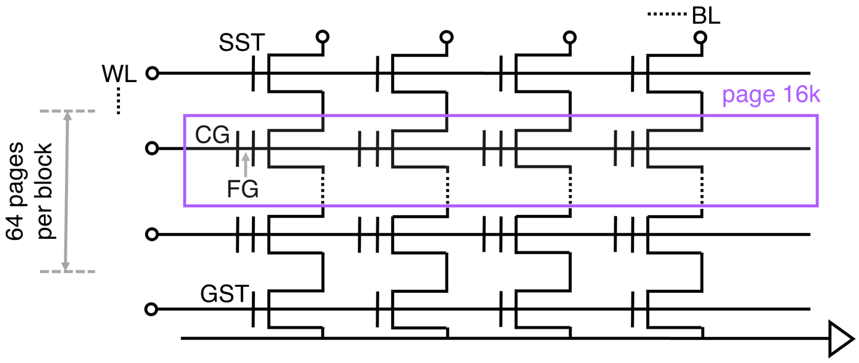


Figure 2.6: NAND-flash architecture for a single block of memory. The circuit schematic symbol for the FG memory cell is a MOSFET with an extra gate. The NAND-flash architecture requires a string-select transistor (SST) and ground select transistor (GST) either side of the bitline (BL) in order to program or erase data onto the cells.

The NAND architecture provides high capacity at low cost. As the cells are connected in long strings across the BLs, interconnects in between cells are not required. Consequently, NAND-flash has a BL and WL scheme that allows the user to select column and row of the memory block. This is, in principle, the minimum number of interconnects needed for a functioning memory array [81]. Strings can be packed together very closely, allowing a per-bit feature size of just 4 F^2 [36]. Moreover, the NAND scaling path has successfully made the transition from planar to 3-D. 3-D scaling will continue in the short term with an increasing number of layers stacked vertically with limited horizontal scaling. The factor that will end 3-D scaling of NAND-flash is not obvious. It is therefore predicted that the cost per bit of the technology will continue to plummet as more bits are crammed onto smaller chip areas [75].

NOR flash

NOR-flash, like NAND, is erased blockwise via FN tunnelling [82]. However, its arrangement (displayed in Fig. 2.7) requires alternative program and read-out procedures, giving rise to some different characteristics compared with the high-density NAND counterpart.

NOR flash is typically programmed bit-wise using HEI. To achieve this, a high voltage is applied to the BL of the target cell, such that electrons conducting in the channel will have sufficient energy to satisfy the inequality given in Equation 2.6 and hop into the FG. This voltage is applied to all cells on the BL. However, HEI cannot occur in absence of a gate voltage, as the cell channel is NORMALLY-OFF. Thus, there are no electrons flowing through the channel at high energy despite the applied BL bias. Simultaneously, a gate voltage is applied to the target cell by accessing the WL; transitioning the cell channel into a conducting state. High energy electrons will only be allowed to flow across the channel of the single target device as HEI requires both a conducting channel and a large electric field across it [82]. This arrangement enables us to program cells individually, suiting this arrangement to low-capacity but high-speed applications, *i.e.* making it unsuitable for NAND. The P/E processes for both NAND and NOR architectures are summarised in Fig. 2.8.

Similarly to NAND-flash, the NOR-flash readout operation requires an intermediate voltage within the threshold window of the device to be applied on the target WL. In this arrangement, we can measure the current of the BL such that if we detect a current, the cell must be conducting, and therefore in state '1'. Likewise, if the BL senses a high resistance, the cell must be in state '0' [82]. This simplistic readout scheme is a direct result of each drain contact of a cell having a path to ground[9].

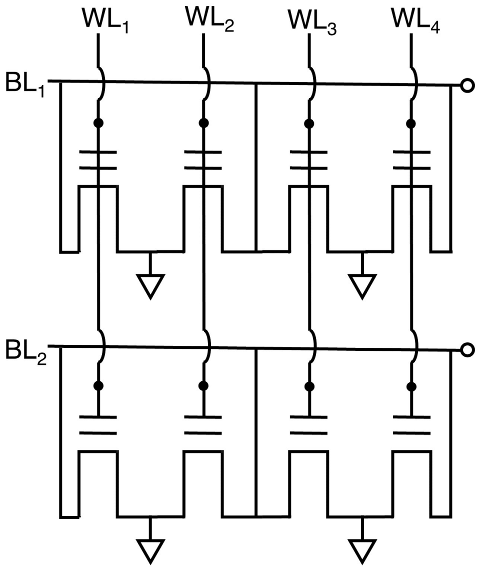


Figure 2.7: Circuit diagram of a NOR-flash memory array. Cells are connected with their drain terminals grounded in pairs for single-bit address.

NOR flash memories are available in capacities up to 2 Gb. Due to the need for separate source and ground interconnects for each bit, NOR-flash is not as scalable or cost effective as NAND-flash for mass storage applications, with its per-bit area of 10 F^2 [36]. However, its arrangement allows for faster access than NAND-flash [36] and the ability to perform bit-wise program cycles. Therefore, it is primarily used in embedded systems where non-volatility is essential (*i.e.* DRAM cannot be used). Examples include reliable code storage (boot, application, OS, and execute-in-place code in an embedded system) and frequently changing small data storage [83].

[9] As opposed to NAND-flash, where the signal must pass through the entire string to access ground.

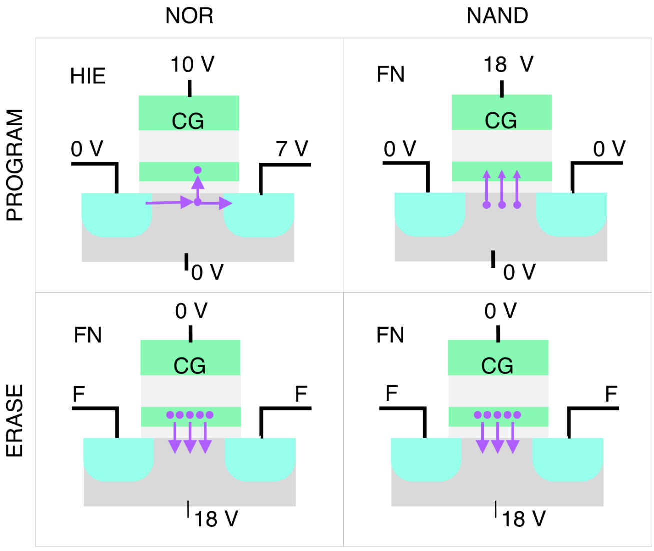


Figure 2.8: Different P/E mechanisms are used depending on array architecture (NOR or NAND), with HEI being used only for NOR-flash. In this example 18 V is required for FN tunnelling and 7 V is required for HEI (across the channel) with the device in the ON state.

2.3 Emerging Memories

2.3.1 Phase change memory

Phase change memory (PCM) exploits a particular behaviour of chalcogenide materials. Chalcogenides can switch phase, from amorphous to crystalline or vice-versa, under heat application. The material remains in this transformed phase once the heat application ceases. Thus, the phase of the material constitutes a non-volatile memory state. This principle is used in compact disk re-writable (CD-RW) memory, where an incident laser changes the optical properties (reflectivity) in a localised area of a chalcogenide glass on the surface of an optical disk. A low-power laser beam is then used to optically read the memory as a digitised signal whilst the disk spins: strong and weak reflected beams correspond to logic 1 and 0 respectively [84]. PCM memory devices operating on similar principles consist of two electrodes sandwiching the chalcogenide (Fig. 2.9). Heat is generated in the chalcogenide phase change material by application of current between the electrodes, resulting in a change in the memory state. The read-out of the memory state is performed by measuring the resistance of the material [85], as the phase change greatly impacts the material conductivity. This results in a large memory window (*i.e.* 0/1 contrast) allowing for large arrays and potentially multi-bit storage. The chalcogenide can be set and reset to form switchable logic states by utilising the temperature and time dependence of the melting and crystallisation properties of the material. Indeed, both high resistance and low resistance states are achieved by electrical heat application with a lower, slower temperature application to achieve a crystalline phase and a fast, high temperature one to melt the chalcogenide to an amorphous phase. [86].

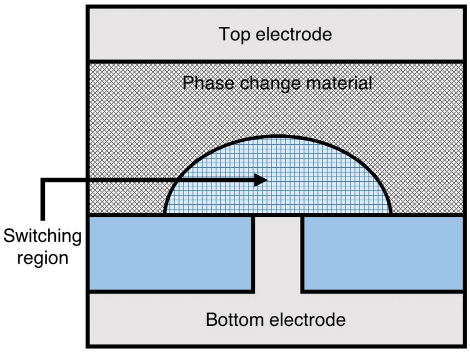


Figure 2.9: Schematic depiction of a PCM cell in which the phase change material switches from crystalline to amorphous in the vicinity of the electrode contact.

PCM was first commercially released in 2015 under the name 3DXPoint. This was the result of a joint development project between Intel and Micron, leading to the memory being available on the consumer market under brand names Optane (Intel) and subsequently QuantX (Micron) by 2017 [87]. In 2016, Micron announced, 'Unlike Phase Change Memory, 3D XPoint technology uses a unique cross point architecture, enabling it to scale in ways that Phase Change Memory has not been able to accomplish [88].' However, it is clear that despite the branding of the cross-point architecture, the operating principle is that of PCM [89]. These technologies are used to bridge the performance gap between DRAM and Flash-based solid state drives, adding another level to the memory hierarchy. The success of PCM is a result of good scalability, 3-D integration ability, fast operation speeds and compatibility with CMOS technology. However, it is not fast enough to replace DRAM, and despite some PCM devices showing endurance capabilities of up to 10^{14} cycles [90], the highest cycling capability of a mass-production one-resistor-per-bit PCM product is just 10^{6} cycles [91]. Moreover, PCM requires a large current, resulting in increased power consumption compared to DRAM [75]. In general, scaling of the PCM cell size is very much limited by the selector devices such as the bipolar junction transistor (BJT) or diode used for cell access [92].

2.3.2 Resistive RAM

Resistive RAM (ReRAM), is a form of non-volatile memory in which the memory state is characterised by the change in resistance of a dielectric material due to a P/E cycle. The operating principles are very similar to PCM[93]. The change in resistance in recent emerging ReRAM technologies relies on the formation and rupture of conductive filaments with a dielectric layer (usually metal-oxide), which is sandwiched between two electrodes. A high applied voltage generates defects in the form of oxygen vacancies in the dielectric layer. Consequently, the vacancies can move under an electric field, thus forming conductive filaments which modulate the resistance of the memory device [94].

For the majority of metal-oxide materials, the oxygen vacancy is a mobile species and therefore constitutes a conduction path for the oxygen ions. The presence or absence of this conduction path defines memory state 1 (low resistance) and 0 (high resistance) respectively. Much like its chalcogenide counterpart, ReRAM offers fast, highly scalable and CMOS compatible memories. There are some success stories; Panasonic™ has effectively used metal-oxide ReRAM as an embedded memory [95]. In specific applications, ReRAM can be very cost effective. However, ReRAM has weaknesses in the 1/0 contrast of its read signal, in which the difference in read current between states is around a factor of ten and typically has a large distribution. Added to this problem is that the cell tends to be non-deterministic, resulting in a large variation in cycle to cycle resistance which further reduces the sensing margin [96].

A ReRAM crossbar array of one-resistor-per-bit (1R) devices would achieve maximum bit density. However, such architectures suffer from sneaking currents through unselected cells that quickly overtake the programming current as soon as the array size grows. Consequently, many ReRAM technologies must use a selection element for each cell, usually in the one-transistor-one-resistor format (1T1R) [97] which requires a larger per-bit area than DRAM of $\sim$ 10-12 F^2 [98].

2.3.3 Conductive-bridge RAM

Conductive-bridge random access memory (CBRAM) is another form of emerging non-volatile memory which relies on changes within the active material to define a memory state, characterised by a resistance modulation. The cell structure of CBRAM is a subset of ReRAM, but here oxygen vacancies are replaced by metal ions. These are formed by fast-diffusive Ag or Cu ions migrating into an electrolyte (which replaces the dielectric layer) [99]. Nanoscale metal filaments can be switched with electrical pulses such that the electrolyte resistance can be modulated to define the memory state. CBRAM cells have one electrode which can easily oxidise into metal ions. Under an electric field, the metal ions can diffuse into the electrolyte, forming conductive filaments which ‘bridge’ a conductive path from one electrode to the other [100].

Similarly to ReRAM and PCM, CBRAM is relatively fast and scalable, and possesses all the advantages of ReRAM whilst eliminating the problematic readout contrast, with a 1/0 resistance modulation of 100-1000$\times$ [101]. CBRAM is emerging onto the memories market, with Adesto™ using this technology to undercut the cost of EEPROM and are recently approaching NOR-Flash array densities [102]. The major disadvantage of CBRAM, however, is its endurance, particularly at elevated temperatures, which is a maximum of 10^5 P/E cycles [103]; an order of magnitude lower than current production PCM.

2.3.4 Ferroelectric RAM

The idea of the Ferroelectric RAM (FeRAM) was first reported in 1974 [104]. Subsequently, a commercial FeRAM was released by 1987 [104]. The cell structure is similar to DRAM, in that each cell consists of a transistor connected to a capacitor. However, the dielectric layer found in the capacitors of traditional DRAM cells is substituted for a ferroelectric material to achieve non-volatility. By applying an electric field, the polarisation of electric dipoles within the ferroelectric layer align with the field direction. Once the electric field is removed, the electrical polarity previously achieved persists due to the material characteristics [105, 106]. The persistence of the electrical polarity in the material comprises the non-volatile memory state in this technology [107]. Thus, the two stable polarisations of the material form binary logic states which can be switched by applying an electrical pulse across the layer with a polarity associated with the ferroelectric polarisation.

FeRAM has a higher endurance than PCM, ReRAM, CBRAM or Flash-memory (10^{13}) [108]. It is also significantly faster than other non-volatile memories. However, the DRAM-inspired structure leads to similar disadvantages. It has a destructive readout procedure which necessitates re-programming any measured data. Moreover, FeRAM faces scaling issues. This is a huge stumbling block for this technology (much more so than DRAM) because it is not straightforward to process ferroelectric materials and electrode materials without causing chemical reactions between the two [108]. For this reason, FeRAM has been stuck at a 15-35 F^2 cell area for many years, more than double that of DRAM; which is already considered fairly low-density [36]. Nevertheless, FeRAM has found some niche applications in smart ID cards and small scale microcontrollers with a small storage capacity [108]. Recently, the discovery of ferroelectricity in doped hafnia provided a breakthrough in scalability of ferroelectric layers. Ferroelectric capacitors were recently reported to successfully scale to 130 nm is a back-end-of-line compatible process, with $<$100 ns switching time. $<$ 4 V switching voltage and good retention properties. Endurance has been confirmed for $> 10^8$ cycles and is statistically predicted to exceed 10^11 when scaled. Although the switching speed and voltage is inferior to more well-known ferroelectric capacitor materials, it is likely that the scaling and fabrication benefits will replace the conventional lead zirconate titanate-based FeRAM [109].

An alternative device construction using similar principles addresses some of the issues with FeRAM. Known as a FeFET, the memory cell resembles a MOSFET in which the oxide material is substituted for a ferroelectric one [92]. Thus, the advantages of ferroelectric memory operation are preserved in a one-transistor-per-bit (1T) architecture. The FeFET is not a new concept, but its development has been impeded mostly by the lack of scalable ferroelectric gate dielectrics. In a similar vein to FeRAM, the discovery of ferroelectricity in doped hafnia reignited interest in this field [110]. 28-nm process compatibility has been demonstrated with a switching speed is as fast as 20 ns, although switching voltage is relatively high ($\sim$ 5 V). Unfortunately, the endurance of the ferroelectric HfO_2 FeFET is currently limited to $\sim 10^{4-6}$ cycles due to parasitic charge-trapping effects which limits its suitability for RAM applications [110, 111].

2.3.5 Magnetoresistive RAM

Magnetoresistive RAM (MRAM) stores data as magnetic domains within a magnetic tunnel junction (MTJ). First introduced in 1989 by IBM [112], the MTJ is formed by two ferromagnetic plates separated by a thin insulating layer (Fig. 2.10). One plate is of a permanent magnetic polarity, whilst the other plate has a magnetisation which can be switched to align with that of an externally applied field. The polarity of magnetisation for the 'free' layer is the basis of the non-volatile memory cell. The electrical resistance of the cell depends of the relative orientation of the magnetisation between the two plates due to tunnel magnetoresistance: the electron-tunnelling probability through the insulating barrier is increased when the two ferromagnetic plates have an aligned polarity. The memory state, therefore, can be read with a simple MTJ cell resistance measurement [113]. Conventionally, parallel magnetisation is interpreted as logic 1 (low-resistance) and antiparallel magnetisation represents logic 0 (high-resistance).

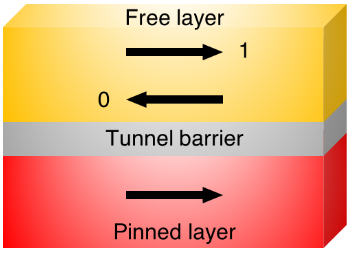


Figure 2.10: Schematic of a magnetic tunnel junction (MTJ) in which arrows are used to represent the magnetic polarisation direction (*i.e.* spin alignment).

The MRAM cell is programmed or erased by flowing electric current on the WL, perpendicular to the cell. Electrons in the free ferromagnet (Fig. 2.10) align to the current direction such that P/E cycles can be performed by alternating the current direction on the WL [114]. The major downfalls of this cell design is near-bit contamination (or disturb) and scalability. The current on the WL can inadvertently align magnetic spins of the free magnets in adjacent cells. This problem is addressed by increasing the separation distance between cells, however bit-density suffers as a consequence [114]. The cell area for MRAM is limited to around 16 F^2, with higher per-bit production costs than SRAM [114].

2.3.6 Spin-torque transfer RAM

Spin-torque transfer RAM (STT-RAM) is the successor to MRAM technology. The two technologies share the use of MTJs to encode data. STT-RAM enables higher densities, low power consumption and reduced cost compared to regular (so-called toggle MRAM) devices. STT-RAM utilises a spin-polarized current to only write to the MTJ cells that need to be switched [108]. Due to its simplicity, a STT-RAM memory cell most commonly follows the one-transistor-one-MTJ (1T1MTJ) model (Fig. 2.11). A MTJ in an MRAM device requires additional bypass lines, separate WLs for write and read and other contacts in order to provide the electrical current to P/E cells. The STT-RAM cell requires just three lines to address the data. The reduction in contact lines greatly improves bit-density [50], reducing cell-area to a size comparable with DRAM [115].

Electrical current is not required to P/E a STT-RAM as only the spin of the electron is being changed. Consequently, spintronic devices can be fabricated in which a spin-transfer can take place without a current present which have the potential for extraordinarily high endurance capabilities. Recent testing has demonstrated that such STT-RAM cells can undergo 10^{15} cycles without any indication of degradation, with an upper endurance limit not yet found [115]. However, STT-RAM currently has only 2-3× read current contrast between its 1/0 states [116]. To produce a high-density, small-feature array a much larger read contrast is required to separate the memory signal from the noise created by multiple sources within the memory core (bit-line, transistor, thermal fluctuations, sense amplifier ect.) [75, 115]. As such, innovative sensing is needed for STT-RAM to overcome this obstacle, and it is a currently the crucial element preventing this technology from large-scale development [75].

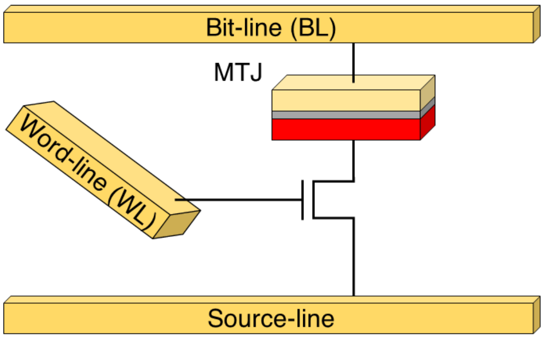


Figure 2.11: Schematic of a STT-RAM cell. The advantage of the STT-RAM cell is that only three interconnects and one transistor are required to utilise the MTJ as a memory-storage device.

2.4 Summary

The performance metrics of each memory technology both in production and in development are presented in Table 2.1. It is generally accepted that the larger the activation energy for the memory state, the more robust (non-volatile) it will be [1, 75]. Currently, there aren't any emerging technologies able to compete with the switching energy and bandwidth offered by DRAM. Although, it is clear based on current emerging memory trends that any competitive fast, low-energy memory should be spintronic or charge-storage based and operate at low-voltages [75]. However, such technologies have failed to provide the scalability, cost effectiveness and robust read signal required to replace conventional DRAM (so far) [36]. The scalability is the most apparent obstacle for most emerging technologies for general working memory applications: most array formats require multiple cell elements per bit. Indeed, recent forecasting suggests that no emerging memory technology will compete with the cell area of DRAM even by 2026 [92].

2.5 ULTRARAM™

Analysts of emerging memories often consider it extremely unlikely that a non-volatile memory would be able to compete with a volatile memory in terms of switching energy (and therefore power consumption). This is essentially a thermodynamic argument whereby a larger energy is required to create a long-lasting state than a shorter-lasting one. A similar argument is often made for switching speed, whereby the length of time it takes to store the state is inherent to its perseverance. In other words, if the state can be programmed very quickly then it must be more fragile than one which required more time and energy to program. Such arguments have led to the notion that a universal memory is unrealistic [1].

ULTRA**RAM**™ seeks to achieve these supposedly contradictory performance characteristics by exploiting quantum mechanical resonant tunnelling to achieve non-volatile logic storage at low switching energies. ULTRA**RAM**™ utilises a FG (similar to flash) to store electrons that define the memory state. However, the single oxide barrier used to transport electrons via FN tunnelling or HEI is substituted for a triple-barrier resonant tunnelling (TBRT) region. Later, the reasoning behind the choice of three barriers and the detailed device physics will be discussed. The materials used for the tunnelling region are InAs and AlSb for wells and barriers respectively. InAs is a high-mobility material with a very low effective mass, making it an ideal choice for high quantum well (QW) confinement energies for memory applications. AlSb provides a 2.1 eV barrier to electrons from its conduction band offset with InAs. Crucially, these III-V materials possess a similar lattice spacing (around 6.1 Å) allowing high quality layer growth.

The quantum properties of the tunnelling structure are carefully engineered (described later in 4.5.1) and result in a device with an extremely long retention time and a required P/E voltage that is an order of magnitude lower than flash memory. As the energy consumption of a FG memory is a square law, this corresponds to a switching energy (per unit area) of at least two orders of magnitude lower than flash memory (assuming similar capacitances). Additionally, as the speed limitation of flash is correlated to the high required voltages (2.2.1), ULTRA**RAM**™ has the potential for high-speed operation, being limited only by capacitances in the array (tunnelling is intrinsically fast). Lastly, the low-energy tunnelling mechanism should not suffer the SILC degradations induced by high-voltages that plague flash technologies. Thus, it is possible that ULTRA**RAM**™'s switching mechanism also offers significant endurance upgrades. The exploitation of the quantum properties available to the InAs/AlSb material system offer a paradigm-shifting approach to charge-based memory technologies. However, this technology comes with its own material challenges which are not well understood. In particular, difficulties arise in wafer-scale fabrication, device-to-device variation, process compatibility and device scalability. However, the obstacles to overcome are thought to be resolvable with continued development and investigation. Next, we will discuss the theoretical and experimental techniques required to design and fabricate ULTRA**RAM**™ memories respectively.

Table 2.1: Benchmarking metrics for memory technologies [1, 36, 75, 92, 117]. Process feature size (F) for size dependent metrics such as switching energy and particle number is assumed to be 20 nm.

Metric	SRAM	DRAM	NOR-flash	3D NAND-flash	PCM	ReRAM	FeRAM	STT-RAM
Cell Area/F^2	120	6	10	<4	4-30	4-12*	15-35	6
Cell elements	6T	1T1C	1T	1T	1T1R	1T1R	1T1C	1T1MTJ
Voltage/V	< 1	< 1	> 10	>10	< 3	< 3	< 3	< 1.5
Switching energy/J	10^{-16}	10^{-15}	10^{-10}	10^{-14}	10^{-10}	10^{-11}	10^{-11}	10^{-13}
Barrier energy/eV	-	0.5	1.6	1.6	2.4	1.4	-	1.5
Retention time	40 ms	60 ms	$>$ 10y	$>$ 10y	$>$ 10y	$>$ 10y	$>$ 10y	$>$ 10y
Endurance	10^{16}	10^{16}	10^{4}	10^{5}	$10^{6} - 10^{13}$	$10^{6} - 10^{12}$	10^{12}	10^{12}
Switching time	1 ns	10 ns	$>$ 100 µs	$>$ 10 µs	100-400 ns	10-100 ns	50 ns	10-50 ns
Bandwidth	-	6.4 GB/s	-	50-250 MB/s	9 MB/s	200 MB/s	1.6 GB/s	2.66 GB/s
Particle	Inverter	e	e	e	Atomic bonds	Oxygen vacancies	Ferroelectric polarisation	Correlated spins
Particle no.	-	10^{4}	10^{4}	10^{3-4}	2×10^{4}	10-1000	-	10^{6}

* $4F^2$ ReRAM are of one-resistor (1R) construction and suffer from reduced performance due to leakage currents.

Chapter 3

Research Methods

This chapter gives an overview of the techniques used to model, grow, characterise, process and measure ULTRA**RAM**™ memory devices and arrays. The material layers were grown by molecular-beam epitaxy (MBE), a portion of which is used for growth characterisation, whilst the majority of material is processed into memories. Memory fabrication requires many processing and characterisation techniques. The details of these are outlined in this chapter; however their specific role in the fabrication process is described in Chapter 5. Detailed simulations of the device physics involved Poisson-Schrödinger and non-equilibrium Green's function with Büttiker scattering calculations combined with a SPICE circuit model. The understanding of device physics gained using these techniques was crucial to the development of the technology.

3.1 Epitaxial Growth Methods

Epitaxy refers to a type of crystal growth (or material deposition) whereby new crystalline layers are deposited on a crystalline substrate which maintain a well-defined orientation with the substrate [130]. The material used in this work, which includes the important heterostructures of the memory technology, is grown using MBE.

MBE is ultra-high vacuum (UHV) evaporation technique. It yields material with impurity levels below ten parts per billion with unparalleled control over the precision with which the composition, layer thickness and doping can be tailored [131]. The end goal of technological development is implementation of the III-V layers on 12" Si wafers via metal-oxide chemical vapour deposition (MOCVD) for commercial reasons. However, MBE serves as a practical growth method for precise layer control in a research setting. Figure 3.1 presents a schematic of a simple MBE process and some of its components. Growing material and dopant sources are positioned around the substrate such that there is a line of sight for material transit. A flux of material (*molecular beam*) is generated by heating the (usually elemental) constituents of the cells (if in the liquid or solid state) or introduced (if gaseous) which in turn causes mass transfer from the effusion cells to the substrate via the vapour phase. This beam travels to the substrate unscathed (*i.e.* without scattering or contacting impurities) due to the UHV. Each source is heated individually for control of material flux and are thermally isolated by liquid nitrogen-filled cryopanels. They each possess a shutter (Fig. 3.1) which controls substrate exposure to each species. The substrate temperature is of crucial importance in surface diffusion, adsorption or desorption of the incoming material atoms. Thus, the substrate temperature is carefully controlled by the operator with a dedicated heater [131].

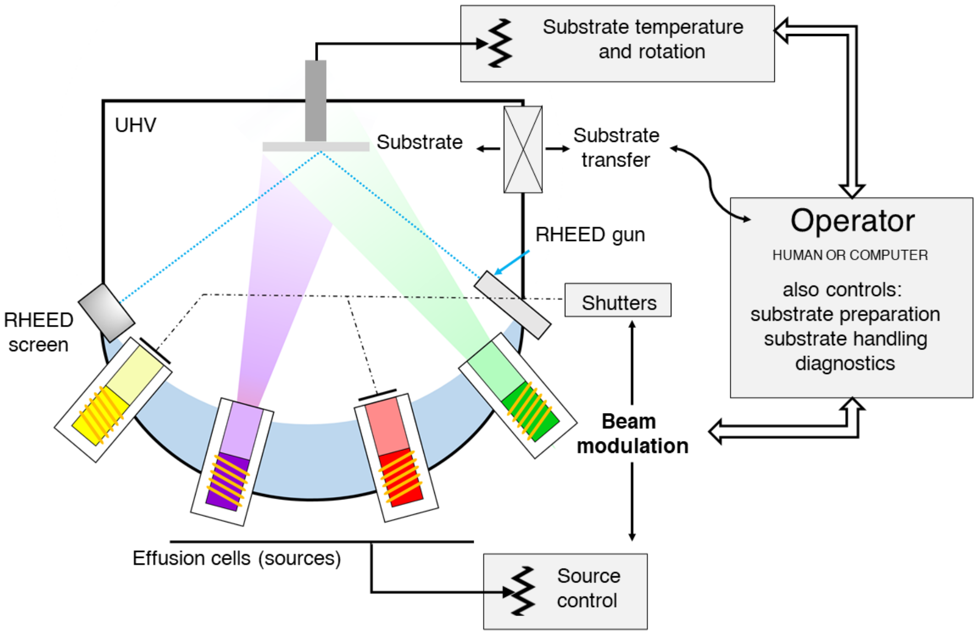


Figure 3.1: Schematic representation of the MBE process and control interface which is controlled by a supervisory operator or computer. The process requires ex situ substrate preparation and wafer introduction procedures.

Reflection high-energy electron diffraction (RHEED) is a technique used for in-situ monitoring of the crystal growth during the MBE process (pictured in Fig. 3.1). RHEED provides information used to verify the cleanliness and crystallinity of the starting surface, as well distinguishing different surface morphologies during growth. The technique is based on the diffraction of an electron beam which strikes the substrate at an angle [132]. A portion of the electrons are diffracted by atoms located at the very surface of the substrate and result in an interference pattern whereby the interference spacing is proportional to the reciprocal lattice vector of the material. The interference pattern presented on the RHEED screen shows this phenomena as streaks or lines. Generally, a streaked pattern indicates a high-quality continuous growth, whilst a spotty pattern indicates a discontinuous growth resulting in poor quality layers [133]. The MBE-grown wafers used to fabricate the ULTRA**RAM**™ prototypes featured in this work were exclusively grown by Dr. Peter D. Hodgson.

3.2 Characterisation Methods

3.2.1 Scanning Electron Microscopy and Energy-dispersive X-ray Spectroscopy

Scanning electron microscopy (SEM) produces images of a specimen using a focused beam of high-energy electrons. It has a higher resolving power than traditional optical microscopy, which is directly proportional to the wavelength of the imaging beam; the wavelength of electrons (2.5 pm at 200 keV) is much smaller than that of optical photons ($\sim$ 500 nm). Practically, the resolution is still limited due to the objective lens system in electron microscopes [118].

The electron beam is generated using a filament or field emission electron gun which is then focused by multiple electromagnetic lenses. The focused beam interacts with the specimen surface, producing Auger electrons, secondary electrons, back-scattered electrons and characteristic X-rays (Fig. 3.2). These electrons provide information about the sample's topography and material composition in the scanned area. The electrons produced by interaction with the specimen are detected as signals and are analysed to obtain an image [119]. Note that this is not a surface-only characterisation technique, as X-rays are produced up to 2 µm below the sample surface.

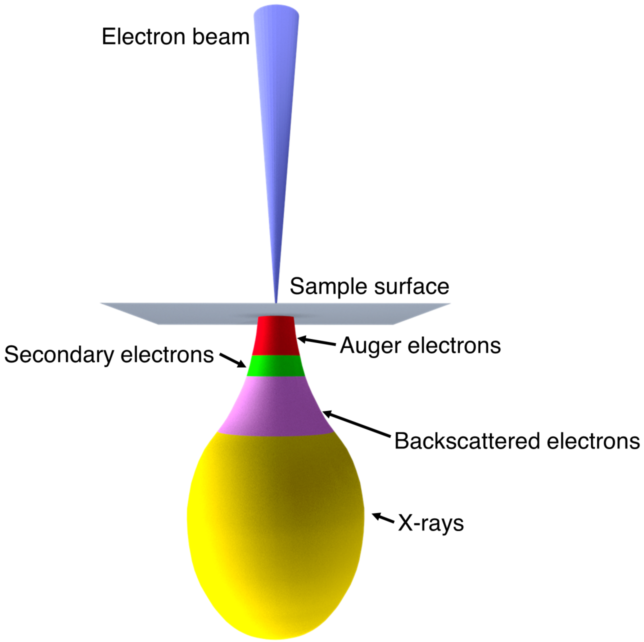


Figure 3.2: Schematic showing the electron beam interaction with a surface with the interaction products, demonstrating the working principle of SEM/EDX. SEM imaging relies on secondary electrons and therefore provides a surface-level image, whereas EDX analyses X-rays for characterisation of buried material layers.

The X-ray emissions produced by the beam-sample interactions allow for energy-dispersive X-ray spectrometry (EDX). This technique analyses X-ray wavelengths and maps them to their corresponding element, from which we can chemically analyse the sample in detail. Coupled with SEM imaging, chemical data can be mapped for an entire surface, a line across the sample or an individual point. Therefore, this method is suitable for chemical species characterisation of complex surfaces [120].

3.2.2 Atomic Force Microscopy

Atomic force microscopy (AFM) is a powerful tool for imaging surfaces at a high-resolution. It was invented by IBM scientists in 1982 [121] and was the precursor to the scanning tunnelling microscope (STM), which later earnt IBM's Gerd Binning and Heinrich Rohrer the Nobel prize for physics (1986) [122]. Image resolutions of the order of fractions of a nanometer are commonplace in AFM systems, although such accuracy comes at the cost of ensuring clean samples, sharp tips and miniaturisation of external vibrations.

The technique relies on the force exerted between the sample surface and the tip to extract information. As the tip approaches the surface, the forces between the tip and sample cause the flexible cantilever holding the tip (Fig. 3.3) to deflect according to Hooke's law. Depending on the circumstances, the nature of the contact force can be mechanical contact, capillary, Van der Walls, chemical bonding, electrostatic or magnetic. The term AFM usually refers to mechanical contact force, where probes using different force detection are given specific names. The position of the tip is precisely located by measuring the cantilever deflection. This is typically achieved by reflecting a laser beam from the cantilever surface (Fig. 3.3) and tracking the beam path using a photodiode detector. The deflection of the cantilever due to the atomic force exerted on the tip results in displacement of the beam path: a measurable quantity. Subsequently, an accurate surface image can be obtained using a feedback loop to control the height of the tip above the sample surface (*i.e.* constant cantilever deflection). Thus, the image is produced from the measurement of the piezoelectric stage as the probe is scanned across the surface [123].

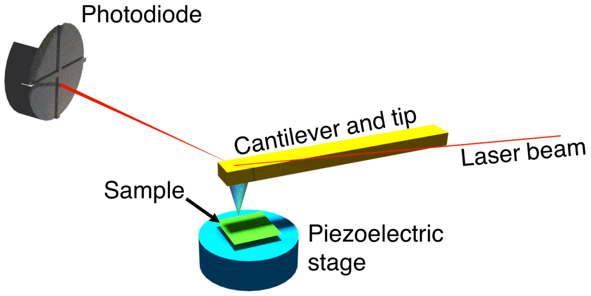


Figure 3.3: Schematic illustration of an AFM system in contact mode operation. The force the sample exerts on the dip bends the cantilever and alters the reflection angle of the laser beam. The piezoelectric stage is adjusted to correct the deflection and map the sample surface.

3.2.3 Beam-exit Ar-ion Cross-sectional Polishing

Beam-exit Ar-ion Cross-sectional Polishing (BEXP) is a material cross-sectioning technique developed by Lancaster University's Prof. Oleg Kolosov, Prof. Manus Hayne, Dr. Ilya Grishin and Dr. Alex Robson. The method uses a modified Ar-ion beam polisher to create a shallow angled (5-10°) slice through the sample [Fig. 3.4**(a)**] to reveal the nanostructures such as quantum wells (QWs) buried beneath the material surface [125]. Conventionally, such material layers are imaged by cross-sectioning perpendicularly to the plane and are then analysed using transmission electron microscopy (TEM). However, the shallow cutting angle produced using the BEXP technique allows access to layers on the material surface. The polishing beam angle spreads the material layers across the surface such that the layer thickness on the surface is magnified which can later be corrected from the known cutting angle.

This technique makes it possible to characterise the layers using surface microscopy techniques (such as AFM, Fig 3.4**(b)**, right) which are generally more cost-effective than TEM. Furthermore, the position of the incident ion beam can be controlled to the micrometer scale [125]. Consequently, the BEXP technique is a powerful, low-cost method of material and process analysis for microelectronic and micro photonic devices as the cross section can be positioned directly on a target device. This technique was carried out on defective memory samples for destructive failure analysis by Dr. Alex Robson.

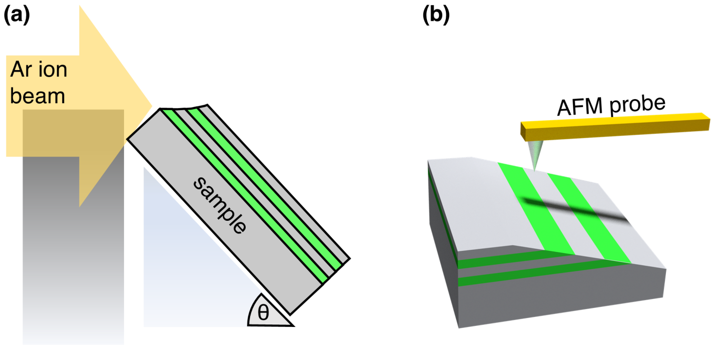


Figure 3.4: Schematic illustration of BEXP of **(a)** angled cross sectioning of a multi-layered sample and **(b)** AFM probing after BEXP.

3.2.4 Mechanical Surface Profiler

A mechanical surface profiler is a characterisation tool used to determine surface features and roughness on a sample. The system features a spherically tipped stylus that moves vertically to contact the surface. The stylus scans laterally thereafter remaining in contact with the sample to map the surface features. The scan speed can be controlled such that a slow scan speed improves measurement accuracy [126].

The stylus position is analysed and the data is available to the user in real time during the scan. Stylus profilers are commonly used because they provide a fast and straightforward method of sample feature measurement to the nanometer scale. However, the measurement is highly dependent on tip size and shape such that the traced profile data is distorted by the stylus. Moreover, the stylus tracking force can harm the sample and measurements are highly sensitive to external vibrations. Thus, the surface profiler is usually reserved for less sensitive samples and measurements [127].

3.2.5 Laser Interferometry

Laser interferometry relies on an external light source (a 670 nm laser diode, in our case), directed onto the the sample surface. The light reflected from the surface is then detected and analysed. The amount of light reflected depends on many factors, including the layer thicknesses of the materials within the sample and their specific optical properties such as refractive index, absorption and reflectance.

In this work, we use a dry etching process (detailed in subsection 3.3.3 of this chapter) to process memory devices. Laser interferometry is an indispensable technique in carrying out this process accurately and reliably. During the etching process, the incident beam is aimed at an area of the sample where material is being removed. The reflectance signal may change when a layer with a different reflectivity is exposed during the etch, resulting in a sudden change in the signal level. This type of measurement is called reflectometry. However, if the underlying layers of the sample are at least partially transparent (as is the case for thin compound semiconductor layers), then the light reflected from the different layers will interfere with each other (Fig. 3.5). This results in constructive or destructive interference depending on the thickness of the layer. Consequently, during the etch process (*i.e.* when the thickness of the material is continuously decreasing), a sinusoidal change in the signal occurs where each period of the oscillation is a 'fringe'. It is possible to count the number of fringes (c) during an etch in order to realise the etching depth. This is given by

$$d = \frac{c\lambda}{2n} \tag{3.1}$$

where d is etch depth, λ is the light wavelength and n refractive index where the interference occurs [128].

Although this gives us a method of depth profiling by counting fringes, this method is insufficient for etching a semiconductor memory structure with multiple thin layers of different materials and thicknesses: the optics are far more complex than those given by the simple equation previously presented. Thus, a transfer-matrix technique is used to simulate the reflectance through a given structure as a function of etch depth (more information on this simulation technique can be found in [129]). Hence, we obtain a simulated etch profile for the exact semiconductor structure which can be mapped to the reflectance data taken alongside the etch process. The result is that the laser interferometry technique, when combined with the simulation results, allows us to control dry etching depth to the nanometer scale [128].

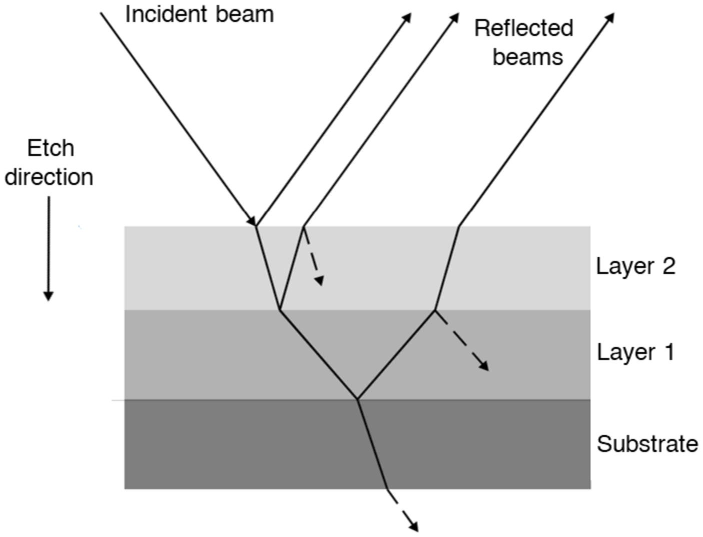


Figure 3.5: The principle of laser reflectance interferometry during an etching process. Note: the incident and reflected beams are all perpendicular to the surface in practice. The angled beams are for illustration purposes only.

3.3 Device Processing Methods

3.3.1 UV photolithography

Photolithography is a process used to transfer a pattern of a thin film onto a substrate, in which later processing steps manipulate the material (usually etching or material deposition), after which the thin film is removed to reveal the desired pattern. The process involves applying a chemical film to the substrate surface which is light-sensitive, known as a photoresist. A photomask is then used to transfer the pattern onto the substrate by selectively allowing light to interact with areas of the resist (Fig. 3.6). The earliest example of a photoresist was invented by Nicephore Niepce in 1826, which used Bitumen of Judea, a natural asphalt, as a coating on a sheet of metal glass or stone. The bitumen became less soluble when exposed to light; unexposed parts of the resist could then be rinsed away (*i.e.* developed) with a suitable solvent to bare the material underneath. This process was used to develop the first ever photograph on a pewter plate, where the developed resist was exposed to an acid dip in order to produce the photograph [134].

Nowadays, the photolithography process remains very similar. The basic patterning process

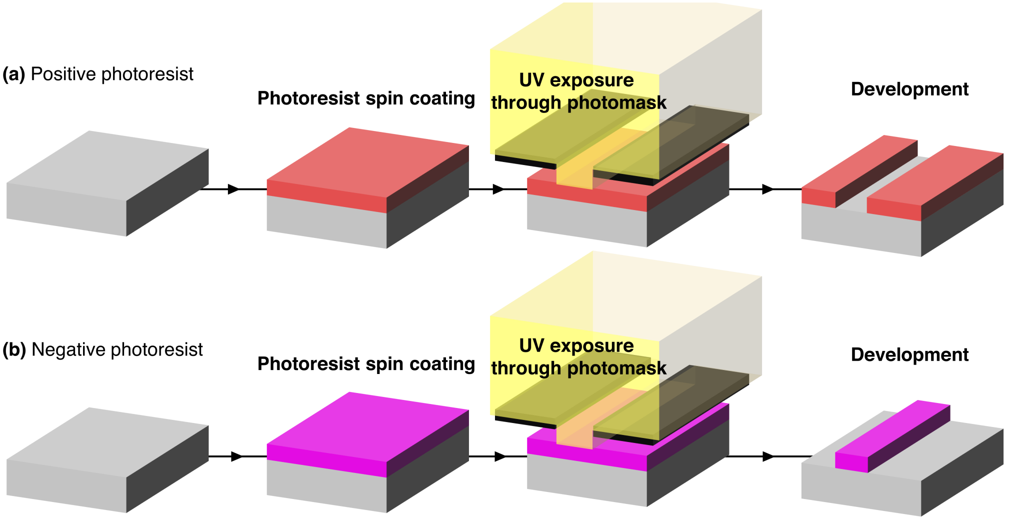


Figure 3.6: Schematic representation of the UV photolithography process for a **(a)** positive photoresist and **(b)** negative photoresist.

for positive and negative photoresists is depicted schematically in Figure 3.6. The long exposure times of Bitumen of Judea are substituted for polymer-based products which are extremely sensitive to UV-radiation, resulting in exposure times of mere seconds and the ability to produce tiny features for microfabrication [135].

Lithographic steps for fabricating microelectronics can be numerous. Microprocessors using deep-UV photolithography processes for sub-µm feature size typically repeat the lithography cycle over 50 times. To achieve this, an alignment tool is required which, in its most basic form, consists of a UV source, a mask holder and a sample holder with the necessary equipment to move the sample into alignment with the mask. In this work, the SUSS MicroTec MJB4 mask aligner uses a high-magnification optical microscope with mechanical adjustment to achieve excellent alignment. The system employs a high-pressure mercury arc lamp to produce near-UV light (365 nm) at a power around 260 W. Soft contact of the sample and photomask is used exclusively in this project, corresponding to a resolution of 2 µm.

3.3.2 Plasma Ashing

Plasma ashing is essentially a cleaning technique used to remove products such as resist residuals from the surface of semiconductor devices during or after the fabrication process. A plasma asher exposes the process gas (mostly oxygen) to the plasma source (high-power radio-frequency waves), ionising the O_2 molecules to produce a monatomic reactive species.

The oxygen plasma reacts with the photoresist residues in which a hydrogen-abstraction step at a hydrocarbon-containing site results in the formation of an alkyl radical. Oxygen addition to the radical sites oxidizes the polymer molecule to generate peroxide and alkoxy radicals. The formation of alkoxy radicals then causes the cleavage of the polymer chain to generate volatile fragments, which ultimately removes the photoresist as an ash to be pumped out of the chamber [136]. Plasma ashing can produce unwanted oxidation reactions that may harm the semiconductor material, especially considering that the chemical ashing process occurs at high temperatures. Consequently, process gas-flow and plasma power have to be carefully adjusted for the specific materials involved.

3.3.3 Inductively-coupled Plasma Etching

Inductively-coupled plasma (ICP) etching is a process technology used to remove material by means of chemical reaction and ion bombardment. Unlike other dry etching techniques, the plasma is generated by electromagnetic induction. A coil (radio-frequency (RF) antenna, Fig. 3.7) encircles the chamber, through which an electric current induces an electromagnetic field within. The electric field accelerates electrons back and forth within the chamber, which is filled with low-pressure gases. This ionises the species in the chamber through collisions, generating a single toroid of high-density plasma [137].

Ions in the plasma constantly bombard the lower electrode in response to the potential difference created by the plasma potential (V_{pp}, Fig. 3.7). A large quantity of the ions cannot react quickly to the RF field but the DC bias set up by the plasma itself (DC self-bias) accelerates them towards the electrode (Fig. 3.7). The ions acquire an average energy (eV) corresponding to a total DC bias which is the sum of DC self-bias and plasma potential. The ion bombardment creates a non-neutral region called the sheath, formed to balance electron and ion losses at the plasma boundary. The ions accelerate through the sheath, reaching the surface at a vertical incidence. This behaviour allows for semiconductor processes involving vertical etching [137]. The independent control of plasma bias (RF coil power) and ion current (forward power) enables high process flexibility [138].

Semiconductor material is removed (etched) through a combination of chemical reactions and ion bombardment. In most cases, ion-bombardment is required to aid the chemical reaction with the surface (*i.e.* etch rate is near-zero in static plasma). Different gases can be used individually or in combination with modified plasma potentials. This choice depends of the physical and chemical properties of the processed material, as well as the desired etch rate and profile of the process. The etching details specific to the material system in this work will be detailed later in Chapter 5 (subsection 5.4.1).

3.3.4 Wet Chemical Etching

Wet etching is a material process which uses liquid chemicals (or etchants) to remove material from the sample. In semiconductor processing, the material is removed in areas which aren't

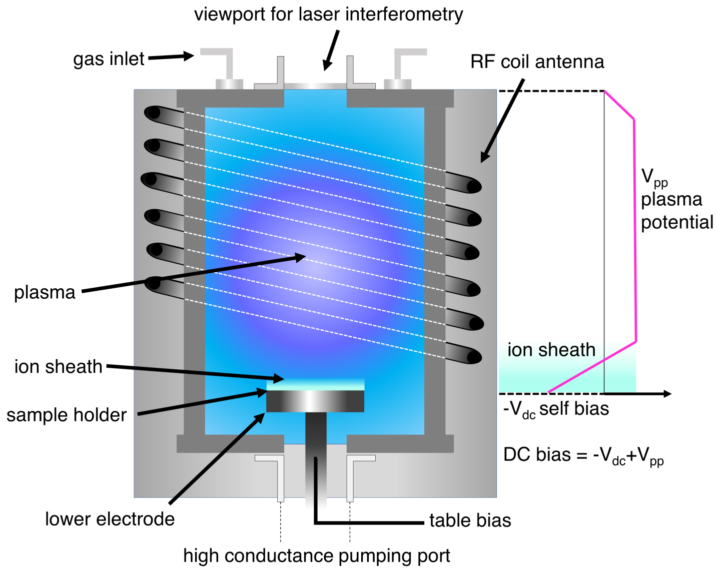


Figure 3.7: Schematic of an ICP etching system alongside the plasma bias through the chamber (pink line).

protected by photoresist as patterned by photolithography. As the process is purely chemical, the etch process is usually isotropic, resulting in lateral etching of the material. Wet etching is a low cost, reliable method and is suited for high production environments with high selectivity[1] in most cases [139]. However, due to its isotropic nature, wet etching is not favoured for devices of small feature sizes where anisotropic dry etching is preferred [140].

3.3.5 Thermal evaporation

Thermal evaporation is a simple method of depositing thin-film material on the surface of a sample. Thermal evaporation requires a vacuum chamber which is evacuated prior and throughout the process. A filament boat or basket is heated up by means of Joule heating by application of current through the boat/basket itself. This boat is filled with the deposition material which, at sufficient temperature, melts and boils to produce a vapour. As presented in Fig. 3.8, the vaporised elements are deposited on the sample surface positioned directly above the evaporation source. The evaporation rate is then monitored using a quartz sensor crystal: a sensor which relies on changes in oscillation frequency due to the added mass from the deposition [142].

High-vacuum levels are required to produce quality film deposition. Moreover, the low-pressure of the chamber during the deposition results in gas particles which travel to the sample with a small probability of collision (*i.e.* their mean free path is larger than the distance between

[1] Selectivity refers to the ratio of etch rate between the selected material etch and the etch rate of an underlying material.

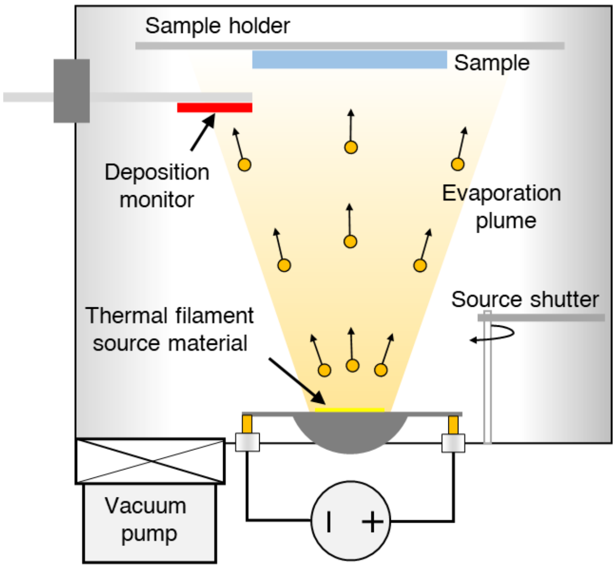


Figure 3.8: Schematic representation of a thermal evaporation system.

source and sample). Consequently, the deposited material arrives at the target surface in an plume of particles with a majority at near-vertical incidence to the surface (Fig. 3.8) [141].

3.3.6 Sputtering

Sputtering is a phenomenon in which particles are expelled from their solid surface due to bombardment by energetic particles of a plasma or gas. This process is an undesirable source of wear for outer-space components [145], but can be useful in the context of thin-film deposition. Sputter deposition utilises bombardment of ions (typically argon) onto a target[2] to release the deposition material into the chamber. A plasma is produced using an DC or RF bias on the target electrode to accelerate the argon ions into collision with the target. RF sources are required if the target material is such that the accelerating potential cannot be applied by DC because the positive charge accumulating on the surface cannot be neutralised [146].

The sputtered atoms from the source material are ejected into the gas phase but are not in thermodynamic equilibrium (as there is no heating involved). Thus, the material tends to deposit on all surfaces in the chamber such that a sample placed within will be coated with a thin-film of the material. Due to the introduction of a precursor gas (argon), the process occurs at high-pressure resulting in short mean-free path for sputtered material. Consequently, the atoms arriving at the sample can be incident at any angle, resulting in an isotropic deposition [141].

[2] A large piece of deposition material mounted on an electrode.

3.3.7 Atomic Layer Deposition

Atomic layer deposition (ALD) is a chemical vapour deposition (CVD) technique capable of producing thin films of a vast array of materials. The technique uses sequential self-limiting reactions to control film thickness at the atomic level, as the name suggests. The resulting film gives exceptional uniformity even on high-aspect ratio features as well as tunable film compositions. The general ALD process is illustrated in Fig. 3.9, consisting of sequential alternating pulses of gaseous precursors which react with the substrate. The deposition of aluminium oxide (Al_2O_3) is a popular choice of dielectric for the ALD technique. As shown in the figure (step (1)), the surface is first exposed to water vapour such that hydroxyl (OH) bonds form on the substrate surface. Typically this occurs naturally due to exposure to atmospheric conditions. Next, a pulse of trimethylaluminium (TMA, $Al(CH_3)_3$) enters the reaction chamber, bonding with the hydroxylated surface to produce a single layer of aluminium atoms each with two methyl (CH_3) groups. Specifically, the half-reaction pathway for step (2) of Fig. 3.9 is;

$$\| - \mathrm{OH} + \mathrm{Al(CH_3)_3} \longrightarrow \| - \mathrm{O} - \mathrm{Al(CH_3)_2} + \mathrm{CH_4} \tag{3.2}$$

where $\|$ denotes the substrate surface. The reaction occurs only when the OH groups are present, thus the reaction is self-limiting, corresponding to the single hydroxylated surface layer [147]. After the TMA reactant is removed, a pulse of water vapour (Fig. 3.9, step (3)) enters the chamber, purging the surface methyl groups according to following reaction [148]:

$$\| - \mathrm{O} - \mathrm{Al(CH_3)_2} + 2\mathrm{H_2O} \longrightarrow \| - \mathrm{O} - \mathrm{Al(OH)_2} + 2\mathrm{CH_4} \tag{3.3}$$

This adds the oxygen to form aluminium oxide layer. The hydroxyl groups regenerate on the surface once more such that steps (3) and (4) can be cycled to build up a high-quality dielectric layer with atomic precision [149]. The total reaction for each cycle is therefore

$$\mathrm{Al(CH_3)_3} + \frac{3}{2}\mathrm{H_2O} \longrightarrow \frac{1}{2}\mathrm{Al_2O_3} + 3\mathrm{CH_4} \tag{3.4}$$

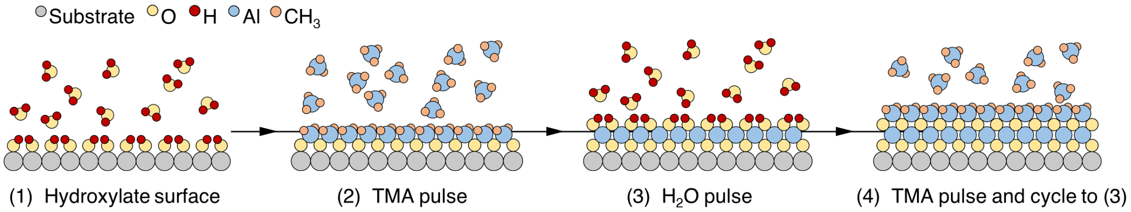


Figure 3.9: Schematic illustration of a two-cycle thermal ALD reaction process for Al_2O_3.

The ALD technique used for depositing Al_2O_3 is popular as it is a high-k dielectric which can be deposited using a simple thermal ALD process with water vapour as the second (purge) precursor. Moreover, the material can be successfully deposited at a wide range of temperatures (33 - 300°C [148, 150]). The simplicity and flexibility of the process makes it an obvious choice for both academic research and industrial applications [151].

It is possible to deposit a wide range of materials using the ALD technique [152]. In the context of microelectronics, Samsung has experimented with ALD to produce high-k and high-quality dielectrics since the late 1990's [153]. This was primarily for the improvement of DRAM capacitors, as previously described in subsection 2.1.2. More recently, the semiconductor industry has transitioned to use high-k dielectrics for the transistor gate stacks in devices. The dielectrics must be highly uniform and pinhole-free to prevent gate oxide leakage currents. Thus, ALD is now favoured for producing non-native oxides on Si. Moreover, the atomic control over the gate dielectric deposition improves process control and solves problems that come about from gate-oxide thickness reductions [152]. In fact, the roll-out of ALD into Intel's mass production line in 2007 was the key factor in the advance from the 65 nm to the 45 nm node technology without creating transistors with significantly higher power consumption [154]. Nowadays, ALD is used in microelectronics production lines by all the major players [155, 156, 157]. Furthermore, ALD is an essential technique in producing high-quality gate-oxides for recent high-aspect ratio transistor designs such as the Fin-FET and tri-gate technologies at the 22 nm node and beyond [158].

3.3.8 Plasma-enhanced chemical vapour deposition

Plasma-enhanced chemical vapour deposition (PECVD) is a material deposition technique first demonstrated by R.C.G Swann at Standard Telecommunication Laboratories, Harlow, Essex in 1959 [159]. The technique shares some experimental similarities with ALD[3]. However, PECVD allows all chemical reactants to enter the pre-vacuumed chamber at once, producing a continuous, isotropic deposition of material onto the substrate (*i.e.* not self-limiting).

Typically, PECVD systems will produce an electric field between electrodes located on opposite sides of the reaction chamber: an RF-energised electrode and a grounded electrode, where the substrate is located on the latter. The field application induces a plasma of ionised reactant gases from the capacitive coupling of the electrodes. The use of plasma enables deposition chemistries at temperatures well below those used in chemical vapour deposition and a film of reaction products is deposited on the heated substrate. Detailed information about this technique can be found in [160, 161].

[3] But precedes the demonstration of ALD by 15 years.

3.4 Transport Measurements

Fig. 3.10 depicts the basic circuit used for electrical measurements of the memory devices of this project. The device symbol resembles that of a Flash-device, as the ULTRA**RAM**™ shares similar operational features. A resonant-tunnelling diode symbol has been added to the conventional FGMOS-symbol, denoting the unique tunnelling mechanism for the technology. Memory readout, program and erase cycles are carried out using a Keithley 2634B dual-channel source-measure unit (SMU) in the configuration shown in Fig. 3.10. Capacitance-voltage (C-V) measurements on larger devices were carried out using an Agilent E4980A precision inductance-capacitance-resistance (LCR) meter. The current measurement shown here (Fig 3.10, labelled *I*) is the sum of the current through the S-D and CG-D under bias; necessitating a low-leakage gate dielectric for S-D current measurement.

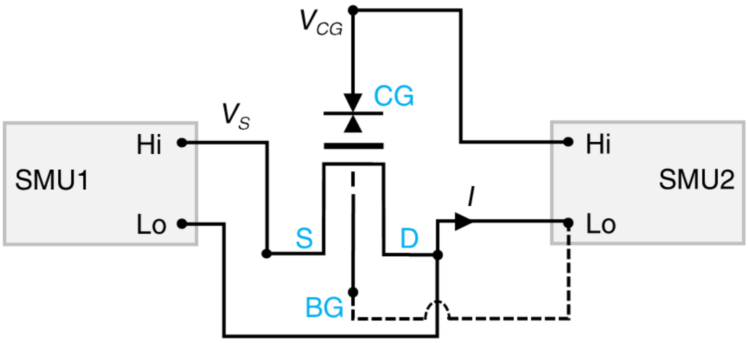


Figure 3.10: Circuit diagram for basic single memory device measurements with a dual-channel SMU. The dashed line for the CG-BG connection indicates that the probe or wire connection must be shifted from the D to the BG for this measurement.

3.4.1 Probe station measurements

A probe station allows electrical measurements to be carried out on the fabricated devices. The sample is placed on a sample holder with the ability to rotate and shift in the direction planar to the surface. Then, careful adjustments are made to position up to four thin, conductive needles (probes) on the areas of the sample where electrical measurement is desired. An optical microscope is located directly above to monitor the position of the probes and sample for accurate placement. The thin probes are attached to conducting wires that are used to send and receive electrical signals to and from the sample.

For single device characterisation, the probes are connected directly to the SMU via triaxial connections as depicted in Fig. 3.10, whereas array measurements are taken through an additional circuit between the probes and the SMU. This circuit is essentially a signal splitter which allows us to switch between devices in the memory arrays with relative ease via manual mechanical switching. In general, the probe station method is used as a fast, straight-forward way of electrical characterisation and the switching box is not often used in conjunction with this method. The input voltage and current sensing is carried out by two source measure units (SMUs) which can simultaneously source and measure DC signals in both sweeping and pulsed operation. These signals are controlled by dedicated LabVIEW programs on a desktop computer; allowing us to sweep or pulse voltages on two terminals simultaneously. This feature is key to characterising the performance of memory devices described in this work. Current sensing within arrays relies on measurements to and from ground due to the nature of the memory architecture (described in section 4.8 in the next chapter).

3.5 Simulation Methods

In the next chapter, the memory concept of this work will be outlined. The device physics of the technology is realised by means of computer modelling. Multiple simulation methods are required to realise the memory performance from our understanding of the quantum transport within the nanostructures. These include two nextnano software packages, nextnano++ and nextnano.MSB as well as a SPICE[4] circuit modelling program.

3.5.1 nextnano++

The nextnano++ software package is specifically designed to simulate semiconductor nanostructures and as such, most of the underlying computational mathematics is hidden from the user. However, it is important to understand the principles which lead to the solutions produced by the software. A detailed explanation is provided in Appendix A.1. Nevertheless, a brief summary is given as follows. The simulation software allows the user to produce an input file (C++ language) which specifies all properties of the device, such as geometry, material composition, strain, doping, grid and contacts. The input file also contains details of the mathematical technique required of the simulation. The program features an in-house Schrödinger-Poisson solver and a material database which includes all III-V compounds and their physical properties. The input is processed and produces solutions which include: band-structure, quantum well energies, wave-functions, electron/hole densities, electrostatic potential and piezoelectric charge [164].

In this work, all Schrödinger-Poisson calculations were carried out at 300 K and gave convergent solutions. The materials database remained unchanged with parameters fixed to well-established experimental values. Consequently, the results obtained from this method are more likely to be replicated by a physical device.

[4] Simulation Program with Integrated Circuit Emphasis.

3.5.2 nextnano MSB

Although nextnano++ includes basic drift diffusion and current continuity equations, these techniques are not sufficient to model resonant tunnelling through a complex nanostructure. As such, nextnano GmbH provides a separate software which uses the non-equilibrium Green's function (NEGF) formalism to calculate the quantum transport through a nanostructure more accurately. The NEGF formalism is the most general and rigorous framework for quantum transport. It ensures that the non-equilibrium carrier distribution in a device is consistently calculated with energy, width and occupancy of its quantum mechanical eigenstates. The program, previously known as nextnano.MSB but has since merged with nextnano.QCL, was developed specifically for calculating current densities in resonant-tunnelling diodes (RTDs) and quantum cascade lasers (QCL) [166]. Again, much of the simulation mathematics is hidden from the user, however an introduction to the NEGF framework is provided in Appendix A.2. The nextnano.MSB technique follows the NEGF framework but sidesteps any self-consistent calculation of lesser self-energies by replacing them by a quasi-equilibrium expression. Doing so is orders of magnitude more efficient than a fully self-consistent nonequilibrium Green's function calculation for realistic devices, yet accurately reproduces the results [165].

The software package accurately takes into account scattering mechanisms in the nanostructure, which include:

- longitudinal polar-optical phonon scattering (polar LO phonon scattering)
- acoustic phonon scattering which includes interface roughness scattering

This is achieved by generalising the so-called Büttiker probe model [166], an explanation of which is provided in Appendix A.3 for the interested reader. Although the NEGF method can be a can be a heavy computational burden, it can accurately replicate the transport characteristics of real-world nanostructures [167, 168].

3.5.3 SPICE

Simulation Program with Integrated Circuit Emphasis (SPICE) is a general-purpose, open-source, analog electronic circuit simulator. It is a program typically used in board-level and integrated-circuit (IC) design to check the integrity of circuit designs and predict circuit behaviour. In this work, it is used to combine the results of the nextnano simulations that describe the quantum operation of the device with a well-established FG memory model. Prior to the addition of the nextnano simulations, the floating-gate SPICE model was confirmed to accurately replicate the expected results for a conventional flash memory device using the Fowler-Nordheim tunnelling equation. The program features include the following simulation analyses [169]:

- AC analysis (linear small-signal frequency domain analysis).

- DC analysis (non-linear quiescent point calculation).
- DC transfer curve analysis (a sequence of non-linear operating points calculated while sweeping an input voltage or current, or a circuit parameter).
- Transient analysis (time-domain large-signal solution of non-linear differential algebraic equations).

LTspice software is used in this work, offering a user-friendly schematic drawing interface which makes circuit simulations straightforward [170].

Chapter 4

ULTRARAM™ Concept and Modelling

The source of ULTRA**RAM**'s unique advantages are derived from it's ability to switch memory logic by the resonant tunnelling of electrons, whilst simultaneously providing a high-energy barrier to retain the memory state after the program (or erase) cycle. Before elaborating on this, we begin by introducing the fundamental physical principles behind the technology.

4.1 Background Theory

This section outlines some important physics relating to the memory concept. The equations used here are simplistic approximations and are provided solely for the reader's understanding of later results.

4.1.1 Quantum well formation

A quantum well (QW) refers to a layer which is sufficiently thin to confine particles (usually electrons or holes) in the direction perpendicular to the layer surface. A quantum well possesses discrete (quantised) energy levels for particle confinement. This can be understood by solving the Schrödinger equation within a potential well. As this topic is discussed in most undergraduate quantum mechanics textbooks [208, 209], we will dispense with the details of the derivation and focus on the results. For a potential well (assuming infinite depth), energy levels are quantised as

$$E_n = E_0 + \frac{n^2\pi^2\hbar^2}{2md} \,, \tag{4.1}$$

where E_0 is the energy at the bottom of the well, d is the QW width, m is the particle mass and n is a positive non-zero integer which numbers the QW states. As a consequence, particles occupying the QW abide by the 2D density of states equation given as

$$g_{2D}(E) = \frac{m}{\pi \hbar^2} , \tag{4.2}$$

where $g_{2D}(E)$ is the density of states (DOS) in terms of energy for one quantised level. Significantly, the DOS does not depend on energy for the 2D case. Thus, as the top of the quantised state energy is reached, there is suddenly a significant number of states available. This results is a staircase-like function where each step corresponds to the next integer n value with a total $g_{2D}(E)$ increase of $\frac{m}{\pi \hbar^2}$, *i.e.*

$$g_{2D}(E) = \frac{m}{\pi \hbar^2} \sum_{i=1}^{n} \mathbf{H}(E - E_i) , \tag{4.3}$$

where $\mathbf{H}(E - E_i)$ is the Heaviside step function and E_i is the i'th energy level. Crucially, no states exist below the E_1 level [210].

Quantum wells can be formed using two semiconductor heterojunctions which are closely separated. Figure 4.1 provides an example using AlSb/InAs heterojunctions in which a thin InAs layer (of thickness d_{InAs}) forms the well and the InAs/AlSb conduction band offsets (ΔE_C) form the potential barriers. These III-V materials are used as examples as they are the materials of choice for the technology. By replacing the parameters of equations 4.1 and 4.2 as:

- $E_0 \longrightarrow E_C^{\text{InAs}}$: the top of the InAs CB
- $m \longrightarrow m_e^*$: the effective mass of electrons in InAs
- $d \longrightarrow d_{\text{InAs}}$: InAs layer thickness

we can deduce the energy levels of the InAs/AlSb QW. The Fermi energy (E_F) describes the highest occupation energy for electrons at absolute zero temperature (Fig. 4.1) and could, in practice, replace E in Equation 4.3 to yield the approximate electron occupation of the QW.

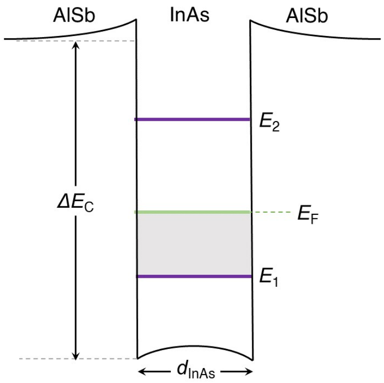


Figure 4.1: Conduction band profile of an AlSb/InAs/AlSb QW.

4.1.2 Resonant tunnelling

Resonant tunnelling is a form of quantum-tunnelling involving two or more barriers. Resonant tunnelling-based devices have been found to exhibit useful features at room temperature and high-bias, unlike most other mesoscopic phenomena that are limited to low temperature response [211]. As such, resonant-tunnelling structures have enabled high-speed, low-power and low-noise devices such as resonant tunnelling diodes (RTDs) and transistors [212, 213, 214].

The double-barrier resonant tunnelling structure features two thin potential barriers with a small separation to form a well. Electrons would easily tunnel through a thin single isolated barrier, however the double barrier structure provides excellent charge-blocking characteristics (*i.e.* high-resistance) at small applied potentials. The current-voltage characteristic of the double-barrier structure can be understood by considering that the region between the two trapping layers acts as a potential well. This well, as described in the previous section, possesses discrete energy levels according to its width [*i.e.* the spacing between the barriers as shown in Fig. 4.2**(a)**]. Assuming that there is only one energy level, E_r, in the energy range of interest [purple line in Fig. 4.2**(a)-(e)**], the system acts as a filter, only allowing carriers with energy E_r to pass through [Fig. 4.2**(c)-(d)**]. Alignments of electron energies with the resonant energy (E_r) are created when a potential is applied to the structure, as this lowers E_r relative to the incident electrons from the emitter. At a threshold energy, the resonant energy falls below the conduction band edge and there is a sharp drop in current [Fig. 4.2**(d)-(e)**] [211].

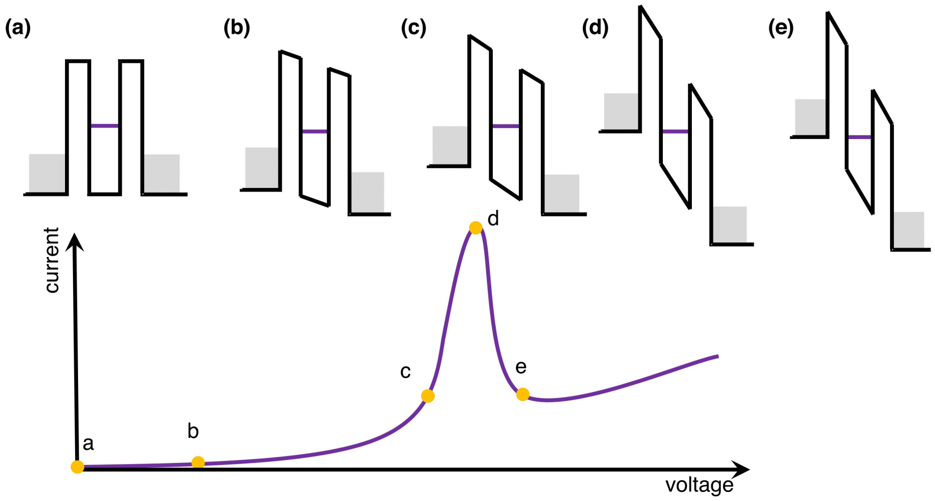


Figure 4.2: Band diagrams of a double-barrier resonant tunnelling structure under increasing potential bias from **(a)-(e)**. The purple line represents the resonant energy state (E_r) and the shaded areas represent the electron population. Each band diagram represents a point on the current-voltage plot, as labelled. This is explained in detail in the text.

The physical reasons for this unique current-voltage relation can be described as follows. As deduced from Equation 4.3 of the previous section, there are no available electron states below E_r. Thus, it is energetically impossible for electrons to hop across the barriers if their energy is less than E_r. This corresponds to tiny current flow should E_r exist above the Fermi level of the emitter [Fig. 4.2**(a)-(b)**]. When sufficient potential bias is applied across the barriers, E_r aligns with electron energies in the emitter. Consequently, electrons of the same energy can hop across the barriers via the quantum well state (E_r). As the electrons are more numerous at the conduction-band edge, the very peak of the current-voltage relation is found when E_r is aligned at this point [Fig. 4.5**(d)**]. After the resonant energy passes the CB edge, E_r lies within the bandgap of the emitter material [Fig. 4.2**(e)**]. There are no carriers in this region, so the tunnelling ceases abruptly [211]. Consequently, there is a sharp decrease in current which is commonly known as a negative differential resistance (NDR) [215].

Resonant tunnelling structures can be fabricated from compound semiconductor heterojunctions. Such devices using InAs/AlSb (InAs wells, AlSb barriers) have demonstrated high peak-to-valley current ratios and large current densities; properties desirable for excellent high-frequency performance [195].

4.2 Resonant Tunnelling Memory concept

The principle of the technology is similar to the floating-gate MOSFET (FGMOSFET) used in flash memory (2.2.1), in that the logic state is defined by confining charges (electrons) to a floating-gate. However, we seek to replace the single tunnelling barrier with a resonant tunnelling structure in order to perform non-volatile switching at low voltages. There are three important considerations when designing a resonant tunnelling structure for memory usage: the position of resonant energies, the number of barriers/wells and the barrier thicknesses/height.

The position of the ground state energy in the QWs of the tunnelling region must be relatively high compared to conventional RTD structures. This is to achieve robust memory retention, as a low-energy ground state will allow thermally excited electrons with sufficient energy to tunnel through the barriers which will result in unintentional electron losses (or gains) in the FG which degrades logic retention. Consequently, the QWs must be designed such that their ground state energies reside significantly above the room temperature Fermi distribution in order to effectively confine FG charges. This is achieved by using very thin layer of a material with a low effective mass (equation 4.1).

Thin tunnelling barriers must be used to achieve high transmission at the resonant energy (high tunnelling current). This allows for low-voltage, high-speed memory switching. The barrier material should also provide a high-energy blockade to prevent electron losses due to thermal emission [284]. When using very thin barriers, the transmission probability at low-energies under zero bias (*i.e.* during retention) becomes significant, which again impacts the non-volatility of the memory logic. Here, a solution is to use three thin barriers to form two QWs with asymmetric ground states. The thin barriers retain electron wavefunction overlap for effective tunnelling where the dual-QWs form a 2D-2D tunnelling process under applied bias. The resulting tunnelling current peak is sharper, and the extra barrier decreases low-energy transmission to retain robust electron storage. The material system required to realise this physical system must be able to produce multiple quantum wells with high confinement energies alongside a complementary high-energy barrier material for robust storage. Moreover, both materials should have the potential to be implemented with high crystalline quality by a precise and scalable method.

4.3 III/V Material System

A III-V material refers to an compounds and alloys made up of elements from columns 13 and 15 of the IUPAC[1] periodic table [171]. These columns are historically known as group III and group V respectively, in accordance the number of valence electrons for an elemental atom. The unique properties of these in binary, ternary or quaternary form have attracted considerable attention as the basis for nanometer-scale optoelectronic and electronic technologies [172].

Figure 4.3 exhibits the bandgap energies and lattice constant of some common III-V semiconductors, along with relevant group IV elemental semiconductors (Si and Ge). The connecting lines indicate the properties of ternary alloys formed by combining the elements in joining binaries at changing compositional ratios. Further, the figure provides information regarding the nature of the bandgap formation. The bandgap is formed by the conduction band-minima and valence band-maxima. Γ, L and X refer to the three valleys of the conduction band [173]. These valleys correspond to specific points of reciprocal-space within the Brillouin zone at which there is high-symmetry [174]. The valley which forms the conduction-band minima is crucial to optical device operation as indirect (*i.e.* not the same *k*-vector) bandgaps greatly suppress electron-hole recombination (photonic emission) [175]. III-V semiconductors crystallise in diamond-like lattice structures similarly to Si or Ge. This is commonly referred to as the zinc blende or ZnS structure [176].

The III-V materials of most importance to this work are InAs, AlSb and GaSb. Highlighted in Figure 4.3 (cyan line), these three semiconductors are approximately lattice-matched around 6.1 Å and occupy a broad range of bandgap energies. The closeness of lattice constant allows for high-quality heterojunctions where lattice strain-related issues are minimised. Moreover, the range of energy bandgaps provides exceptional flexibility in semiconductor device engineering. As such, InAs/AlSb/GaSb heterostructures have proven to be an excellent choice for many electronic and optoelectronic devices; including high electron-mobility transistors (HEMTs) [177], field effect transistors [178, 179], infrared detectors [180, 181] and semiconductor lasers [182, 183].

[1] International Union of Pure and Applied Chemistry.

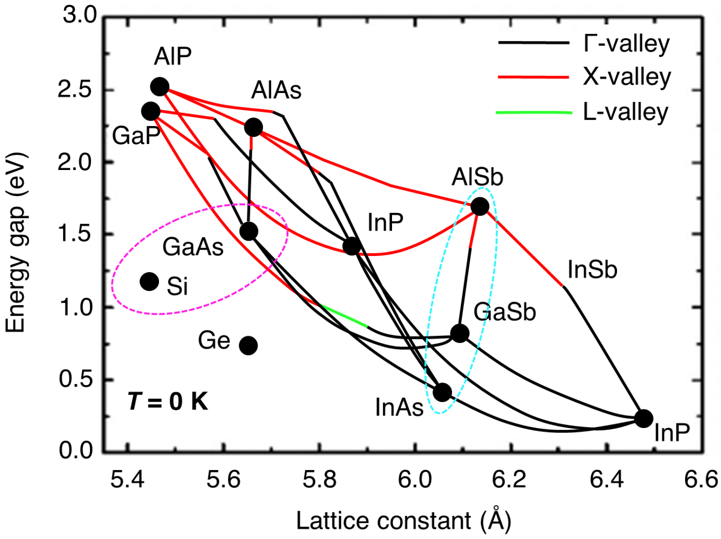


Figure 4.3: Energy bandgap and lattice constants of some binary III-V semiconductor materials. Connecting lines indicate the ternary alloys of the connected compounds at various ratios of corresponding binary materials. Elemental group IV semiconductors (Si and Ge) are included for later reference. The cyan oval indicates the 6.1 Å semiconductor family, whilst the magenta oval points out the substrate materials used in this work. Adapted from [197].

In theorising and fabricating a semiconductor memory suitable for mass implementation, one must consider the production costs of such undertakings. GaSb is available at a maximum wafer diameter of 4". In order to be cost competitive, the memory concept is first developed on GaAs: available on 8" wafers for less than one third of the price of GaSb [184]. More importantly, larger wafers greatly reduce processing costs as the number of chips which can be processed simultaneously increases. Consequently, the ultimate platform would be 12" silicon, where ULTRA**RAM** ™ could fit seamlessly into mass production lines in major silicon fabrication plants. The details of implementation on GaAs and Si will be discussed in due course, in which the jump in lattice constant (Fig. 4.3) is carefully considered. However, it is important to note that the choice of GaAs for the substrate within the theoretical modelling section of this chapter is a practical one and does not effect the overall results of the simulations where lattice strain is minimised.

4.3.1 InAs/AlSb Heterojunction

Heterojunctions are formed when two layers of dissimilar crystalline semiconductor material interface. The semiconducting materials have unequal bandgap energies such that conduction and valence band (VB) energy band alignments can be engineered for specific electronic or optoelectronic device applications [185]. These band-energy alignments fall under three distinct categories: straddling-gap (type I), staggered gap (type II) or broken gap (type III) [186]. The InAs/AlSb interface is an example of a type-II heterojunction (Fig. 4.4).

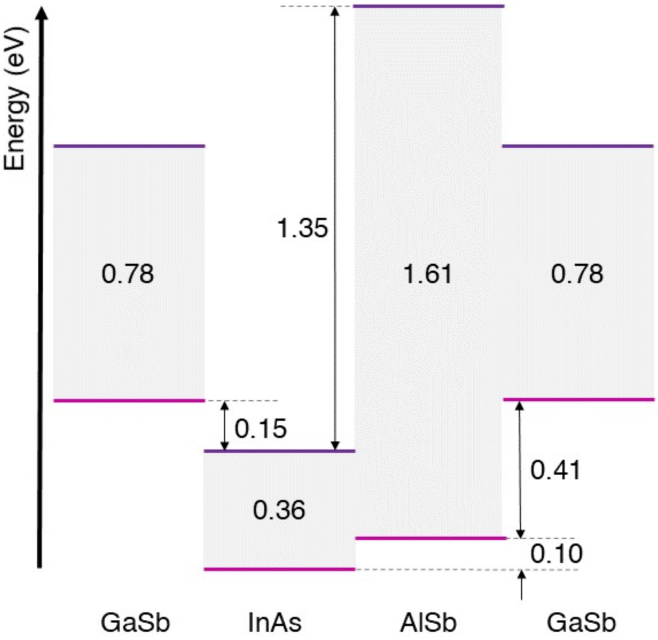


Figure 4.4: Energy band line-ups for the 6.1 Å family of materials. Note that the InAs/AlSb band offset is 1.35 eV here [187].

InAs is a compound semiconductor formed from indium (group-III) and arsenic (group-V) with a lattice constant around 6.06 Å and a narrow room temperature bandgap energy of 0.36 eV. The renowned qualities of InAs are its low electron effective mass (m_e^*) and high electron mobility (μ). These essential properties have made InAs an attractive channel material in compound-semiconductor CMOS electronics [188, 189, 190]. In the context of memory applications, InAs is a superior choice for a QW material, as it's extraordinarily low effective mass allows for high confinement energies which are essential for memory retention. Indeed, achieving a QW ground state energy as large as that of the InAs/AlSb QWs in this work using the (more common) GaAs/AlGaAs system would require a QW width of less than one lattice constant (equation 4.1).

AlSb is an indirect-gap semiconductor (L-valley, Fig. 4.5) formed from aluminium (group-III) and antimony (group-V) with a lattice constant around 6.14 Å. AlSb has received considerable attention as a barrier material of high mobility electronics: including HEMTs, quantum cascade lasers (QCLs) and resonant tunnelling diodes (RTDs) [191]. AlSb is favoured due to its band-alignments within the 6.1 Å family, with a large bandgap compared to InAs and GaSb (Fig. 4.4). The main disadvantage of this material is that it is highly prone to oxidation [192].

Figure 4.4 shows the band lineups in the 6.1 Å family of semiconductors. The InAs/AlSb conduction band offset is stated as 1.35 eV here. However, the underlying physics of this heterojunction is more complex. The bandgap energy picture of semiconductor physics is a simplification of the overall band diagram; the bandgap is the gap between conduction band minima and VB maxima of the full band diagram (Fig 4.5). For direct-gap[2] material heterojunctions, the minima-maxima bandgap model is sufficient. Indirect bandgap semiconductors such as AlSb must consider crystal momentum wavevector (k). Figure 4.5 displays the full band-diagram for **(a)** InAs and **(b)** AlSb as a function of crystal momentum (k). The 1.35 eV InAs/AlSb band offset is anchored from the conduction band minima at the L-valley of the AlSb band diagram. However, this is separated from the conduction band minima of the InAs Γ-valley in k-space (L-Γ), such that a charge carrier must interact with phonon(s) of 0.73 eV to traverse the indirect gap [193]. Thus, any interaction across the heterojunction is much more likely to occur through the Γ-valley minima of the AlSb where high-energy phonon interaction(s) are not required. Consequently, the conduction band-offset of the InAs/AlSb heterojunction forms a 2.1 eV energy barrier for electron transport through it (Fig. 4.5, from E_Γ = 2.22 eV) [194, 195]. This band-offset energy is extraordinarily large and forms the basis for charge storage within the memory devices. In fact, the InAs/AlSb material system is suggested to be the ultimate choice for compound semiconductor memories [196].

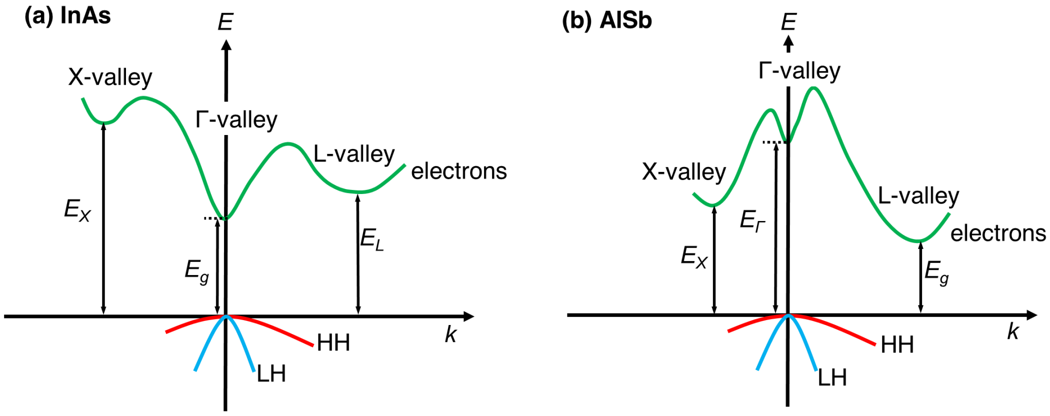


Figure 4.5: Bandstructure of **(a)** InAs and **(b)** AlSb as a function of wavevector (k). Heavy hole (HH) and light hole (LH) valence bands are typical schematic depictions, whereas conduction bands are simulation data.

4.3.2 InAs/GaSb Heterojunction

GaSb has a direct bandgap (Γ) of energy $E_g = 0.73$ eV at room temperature [197] and a lattice constant of 6.10 Å. When GaSb interfaces with InAs, the conduction band offset is around 0.9 eV and VB offset is approximately 0.5 eV. Consequently, the InAs CB is located at a lower energy than the GaSb VB resulting in a type-III broken gap heterojunction (Fig. 4.4) [198]. The atypical band alignment allows electrons to travel from the InAs conduction band to the available states in the GaSb VB. As a result, the InAs/GaSb heterojunction demonstrates semi-metallic properties.

[2] The conduction-band minima and VB maxima occur at the same crystal momentum (k).

The unusual properties of the InAs/GaSb heterojunction has led to numerous investigations of quantum well (QW) and superlattice (SL) structures with specific applications in mid and long-wave optelectronic devices such as infrared lasers, detectors and photodiodes [199, 200, 201]. In electronics, this heterostructure has yielded promising results as a tunnel-FET (TFET) structure [202]. Additionally, the unique bandstructure of this system has been exploited to fabricate dopant-free n-MOS and p-MOS transistors by manipulating carrier concentrations across the heterojunction under electric fields [203, 204].

4.3.3 InAs/Al_2O_3 heterojunction

The term heterojunction does not exclusively refer to semiconductor-semiconductor interfaces. In this work, Al_2O_3 is used as an insulating gate dielectric to form an InAs/Al_2O_3 semiconductor-insulator heterojunction. The material can be deposited ‘gate-last’ such that it must be added after several fabrication steps. When deposited by thermal ALD on InAs, the conduction band and VB offsets are 3.1± 0.1 eV and 2.5± 0.1 eV respectively [205]. Therefore, Al_2O_3 provides sufficient blocking-barrier properties to be utilised as a gate-dielectric on InAs. In fact, this has been used successfully in [202, 206] to this end.

The InAs/Al_2O_3 heterojunction is extremely sensitive to the deposition technique and surface preparation of the sample prior to heterojunction formation. In general, InAs interfaces with metals and dielectric materials cause Fermi level pinning, whereby the Fermi level is pinned above the InAs CB [207]. This point will be elaborated on in due course (5.3.4), however for the purposes of this chapter it is important to note that the InAs/Al_2O_3 interface produces an unpinned Fermi level when formed via ALD and that the InAs Fermi level is pinned above the CB for all other interfaces (metals, oxides and air) apart from those heterojunctions formed with GaSb and AlSb (MBE-grown crystal).

4.4 ULTRARAM™ Concept

4.4.1 Program and erase via resonant tunnelling

The fundamental principle on which ULTRA**RAM**™ is based is of a floating gate (FG) memory. Similarly to FGMOSFETs used in Flash memory (2.2.1), the logic states (*i.e.* 0 or 1) of the memory are defined by the presence or absence of charge within the FG. Electrons are transported into and out of the FG via a tunnelling mechanism under applied electric field, after which they remain trapped there by potential barriers. ULTRA**RAM**™ implements a triple-barrier resonant tunnelling (TBRT) structure utilising the extraordinarily large CB offsets of the InAs/AlSb heterojunction (2.1 eV) to achieve electron barriers akin to those of dielectrics and hence non-volatility. However, the triple-barrier resonant tunnelling structure, as pictured schematically in Fig. 4.6, allows electrons to tunnel in and out of the FG under low bias. This resolves the paradox of universal memory, as the tunnelling structure provides a high-energy barrier when there is no bias applied, but allows resonant-tunnelling (*i.e.* transparent barriers) at program/erase (P/E) voltages of around 2.5 V, approximately 10 times lower than Flash [75].

The TBRT structure of the memory is the source of its incredible performance characteristics, which will be discussed later in Section 4.7. The materials and choices of layer thicknesses for the TBRT region are consistent throughout all evolving iterations of the technology, including those of the initial breakthrough single memory cells [216]. Changes to the overall design were motivated by attempting to improve endurance, readout contrast, architecture compatibility and silicon integration, all of which will be discussed in due course.

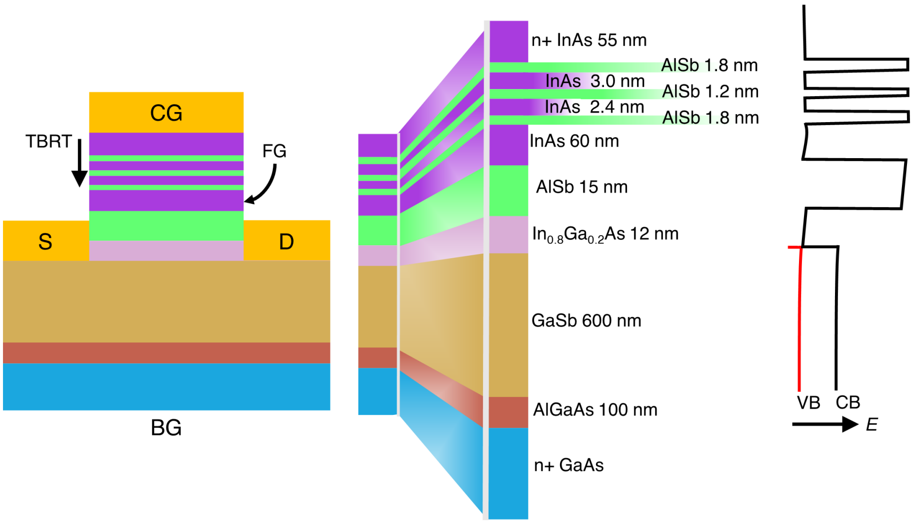


Figure 4.6: Schematic of an ULTRA**RAM**™ prototype device structure, which is the subject of the first device modelling work. CB and VB of the structure presented alongside are 300 K simulations which have been adapted to align with the schematic layers (not to scale).

4.4.2 Memory readout

As a FG memory, ULTRA**RAM**™'s readout mechanism (*i.e.* logic state determination) is carried out by detecting the presence or absence of charge in the FG. Much like flash memory, ULTRA**RAM**™ implements a channel positioned underneath the FG with source (S) and drain (D) contacts attached. The amount of charge in the FG above modulates channel conductivity such that the memory logic can be inferred from a measurement of conductivity through the channel.

The breakthrough prototype devices (*v1.0*) [216] utilised a n-type InAs channel of 60 nm thickness. This channel is highly conductive with the FG empty (*i.e.* NORMALLY-ON[3], memory state 1). With the FG filled with electrons (0) the electrons in the FG deplete the carriers in underlying n-InAs channel and therefore reduce the channel conductivity [216]. The resulting modulation in conductivity is significant and the 0 and 1 states can be measured as a change in current reading across the channel (S-D). However, a commercial, high-density memory requires potentially 1000's of cells to be connected in series to form a bit-line (BL). For this to be made possible, a seismic improvement in readout contrast (0/1) is of paramount importance [75]. Fortunately, the insufficient read contrast is not an indication of logic state weakness, but rather due to the simplicity of the channel construction. For compatibility with compact (1T) architectures, a NORMALLY-OFF[4] state would be preferred, in which a threshold voltage (V_T) is formed by an applied CG bias, and a shift in threshold voltage (ΔV_T, equation 2.3) is caused by the screening effect of the FG electrons to define the memory state (much like Flash, see 2.2.1).

For silicon-based memories, a threshold voltage in achieved through lateral doping of the channel to form p-n junctions. Although this process can be successfully carried out on III-V materials such an InAs [217], doing so significantly adds to the cost and complexity of device processing. A newly proposed readout mechanism achieves the desired channel properties without lateral doping and with only small changes to the previous layer structure: The bulk-InAs channel is substituted for a 12 nm $In_{0.8}Ga_{0.2}As$ quantum well (QW_{CH}), depicted in the schematic shown in Figure 4.6. The introduction of gallium into the InAs alloy, coupled with the QW state formed from the thin channel, raises the minimum energy level of the channel (QW_{CH}) CB above the Fermi energy at zero bias. Consequently, the (now intrinsic) InGaAs channel has a CB void of electrons and is therefore highly resistive (*i.e.* NORMALLY-OFF). If a bias is applied across the structure from CG to BG (Fig. 4.6), the bandstructure shifts such that the channel fills with electrons and causes a sudden increase in conductivity. It is this process that constitutes the current modulation and forms a threshold voltage (V_T) corresponding to the energy of the channel QW energy (QW_{CH}) in relation to the GaSb VB. Detailed modelling of the channel allows us to understand and predict its behaviour and optimise thicknesses and composition of layers (Section 4.6).

4.5 Resonant Tunnelling Program/Erase Simulation using nextnano

4.5.1 Resonant tunnelling simulation as a RTD

nextnano.MSB is a program specifically developed to model QCLs and RTDs as previously detailed in subsection 3.5.2. This software package is used to model the resonant tunnelling program/erase (P/E) cycles. To do this, only the triple-barrier tunnelling region is modelled such that the simulation probes are positioned on the CG and FG of the device (Fig. 4.7). Consequently, the

[3] The channel is highly conductive under zero bias and can be switched to higher-resistance under applied bias.

[4] The channel has a high resistance under zero bias and can be switched to a conducting-state under applied bias.

Table 4.1: nextnano.MSB material parameters

Parameter	InAs	AlSb
VB offset (eV)	1.390	1.385
Band-edge gap (eV)	0.417	2.386
Band-edge α (eVK^{-1})	0.276 $\times 10^{-3}$	0.42 $\times 10^{-3}$
Band-edge β (K)	93	140
Effective mass (m_0)	0.026	0.14
Static dielectric constant	15.15	12.04
Optic dielectric constant	12.25	10.24
Deformation potential (eV)	-6.66	-8.12
Material density (kgm^{-3})	5.61 $\times 10^3$	4.26 $\times 10^3$
LO phonon energy (meV)	30	42
LO phonon width (meV)	3	3

simulation is arranged as an RTD whereby the current density modelling describes the movement of electrons into and out of the FG (P/E). The simulation parameters used to model the device physics are provided in Table 4.1 and are fixed to experimentally observed constants [166, 197] for the material layers of the TBRT region.

The triple barrier construction of the tunnelling region forms two QWs within the structure (labelled QW$_1$ and QW$_2$, Fig. 4.7), causing electrons to be confined to distinct energy levels. Two quantum wells are required to produce a sufficiently thick barrier to prevent leakage via conventional tunnelling (*i.e.* not through a resonant state), whilst simultaneously using thin QWs raises the confined states to produce well-defined RT peaks. Moreover, the well thicknesses are sufficiently dissimilar to prevent energy-state alignment between the two wells, which would otherwise reduce the electron blocking capability of the central barrier [218]. The TBRT design used in this model is studied in detail and compared against alternate options later in Chapter 7.

Figure 4.8 shows the simulation results for the TBRT region as an RTD when scanning through voltages of positive and negative polarity. The colour scale DOS clearly indicates the confinement of available states in the QWs. Applying a voltage across the tunnelling junction tilts the conduction band such that the energy levels relative to the energy of incident electrons (emitter) changes. In the case of this structure, the electrons outside the tunnelling junction are in a quasi-bound state due to the formation of a triangular-shaped well from the applied voltage [218]. This is shown by the colour scale for the density of states (DOS) for the write process displayed in Fig. 4.8**(c)** and **(d)**. In these figures, the conduction band is at a gradient due to an applied voltage at the CG of the device.

Similarly to a conventional double barrier RTD, resonant tunnelling occurs when the energy state of the emitter aligns to the resonant energy level formed in the quantum well(s). The DOS

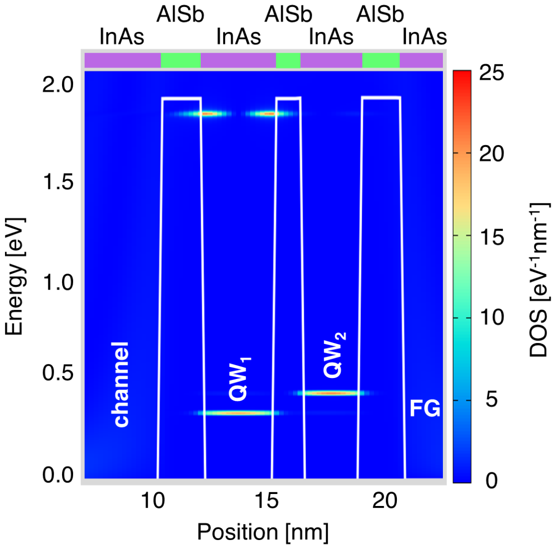


Figure 4.7: Density of states (DOS) simulation for electrons in the TBRT region under zero bias. The white line is the CB where offsets are formed from InAs/AlSb heterojunctions. The DOS are shown by colour scale.

of the energy of QW$_1$ is spatially present in QW$_2$ and vice versa, which can be seen faintly in the DOS plots in Fig. 4.8**(a)-(d)**. This is a result of electron wavefunction overlap across wells due to the thin central barrier. Consequently, the triple-barrier structure has two resonant energies associated with QW$_1$ and QW$_2$ and we predict a double-peaked current density relation. This is confirmed by the current-density simulations (Fig. 4.8**(e)**, black line for program cycle). The current-density plot [Fig. 4.8**(e)**] for program and erase cycles, in which the program cycle refers to moving electrons into the FG, has peaks which correspond to the energy alignments of QW$_1$ and QW$_2$ which are a result of coherent resonant tunnelling.

Modelling of the TBRT region as a RTD has provided us with a thorough understanding of the quantum transport of the device. Moreover, the simulation has provided a current-density vs applied bias (1-dimensional) plot for P/E cycles. Next, this data will be used to further our understanding of the device physics as we endeavour to form a more complete model of the memory cells.

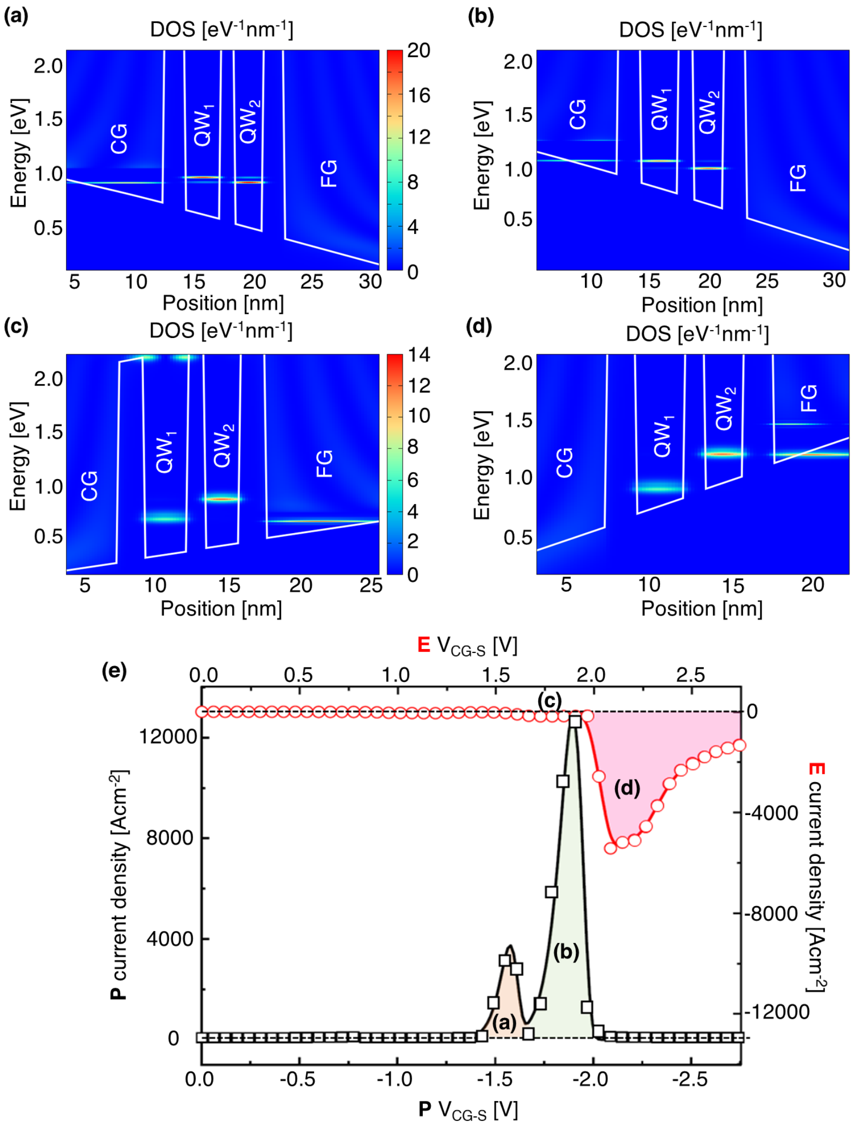


Figure 4.8: Simulation results (300 K) for the tunnelling region of the device. **(a)-(d)**, QW energy levels for the structure are shown where the colour scale indicates the electron density of states (DOS). **(a)** -1.6 V CG bias for the write cycle. **(b)** -1.9 V CG bias for the write cycle. **(c)** +1.7 V CG bias for the erase cycle. **(d)** +2.1 V CG bias for the erase cycle. **(e)** Current density to CG-channel voltage relation for the write (black) and erase (red) cycles. Labels **(a)**, **(b)**, **(c)** and **(d)** correspond to the simulation results in the respective parts of the figure.

4.5.2 Gate stack corrections

The next step in extending this model is to factor in the gate stack of the device. Here, we aim to convert the voltage across the TBRT region in the nextnano.MSB simulation (V_{TBRT}) into the voltage applied to the device terminals. This allows us to investigate the P/E cycles for the modelled memory device.

The gate stack is simulated for the device in 1-D at 300 K using nextnano++ (detailed in 3.5.1). These calculations include the band-bending effects of the doping densities within. Charge neutral contacts[5] are added within the simulation space positioned at the CG, and the $In_{0.8}Ga_{0.2}As$ channel regions of the device. A voltage sweep is applied and the bandstructure, energy eigenvalues, charge densities and electron WFs and probabilities are calculated for each voltage interval and matched to the calculations under V_{TBRT} to transform tunnelling simulation voltages into voltages applied at the device terminals.

Conventionally, the gate stack is considered using a capacitative coupling approximation; the gate stack is treated as a combination of capacitors and wires [62, 219]. However, this method neglects the properties of the semiconducting materials, omitting band-bending effects, dopants, strain and minor carriers (holes) and are therefore inferior to using Poisson-Schrödinger simulations. Moreover, energy eigenvalues of the QW states (QW_1, QW_2) were compared between the MSB and nextnano++ equivalent voltages (*i.e.* similar CB slope across the TBRT region). The energy eigenvalues were extremely similar, from which we conclude that the resonant energy alignments (thus coherent-resonant tunnelling) will occur at V_{CG-S} voltages matching V_{TBRT} values for the TBRT current-density peaks. It is important to note that the previous results [Fig. 4.8**(e)**] already include this voltage correction to the gate stack.

4.6 High-Contrast Readout Simulation using nextnano and SPICE

Readout is extremely important to memory operation: it is a major stumbling block for some promising emerging memory technologies (such as ReRAM and MRAM). The readout measurement procedure should maximise the contrast between the logic (0/1) memory states, known as readout contrast. A poor readout contrast limits the capacity of the memory array and the architectural flexibility in the memory design. Fundamentally, the number of memory cells (bits) that can be placed in-line (*i.e.* on the same wire; the bit-line) is limited by the ability to distinguish between the logic measurement and the leakage of cells sharing the same bit-line. As a consequence, the number of cells on the bit-line should not exceed the readout-contrast of the memory [75]. Additionally, a memory cell which has a high-resistance when no power is applied (*i.e.* NORMALLY-OFF) reduces power consumption and allows access to individual memory cells whilst the surrounding bits remain electrically isolated. This reduces the disturb rate and allows for flexibility in designing the memory architecture. FGMOSFETs found in Flash typically use NORMALLY-OFF channel designs with use of a threshold voltage (see subsection 2.2.1) achieved through lateral doping to produce outstanding 0/1 contrast. Despite this, flash memory cores rarely exceed 64-pages[6], whereby

[5] The Dirichlet value for the potential within the contact is determined by requiring local charge neutrality for each grid point of the contact.

[6] However page length is 16 kbits, producing large memory arrays (2.2.1).

readout is hindered by parasitic capacitances [220].

In this work, we aim to replicate the advantages of the flash readout mechanism to produce a NORMALLY-OFF channel with a well-defined threshold voltage (V_T) for logic state determination. By combining the type-III (staggered-gap) band offset of the In(Ga)As/GaSb with QW confinement energies, it is possible to achieve these properties without the use of lateral doping. This simplifies the fabrication process and would (hypothetically) greatly decrease production costs should ULTRA**RAM**™ reach the commercialisation phase. First, we consider the bulk properties of the InAs/GaSb heterojunction. Schrödinger-Poisson simulations (nextnano++) are depicted in Fig. 4.9. The alignment of the bandstructure results in a gapless junction where the InAs CB lies below the GaSb VB. As a result, the electron population of the GaSb VB is able to proceed into the InAs CB, filling the CB with electrons to the energy of the GaSb VB maxima. This phemonemon is clearly demonstrated by the simulation results presented in Figure 4.9**(a)**, where there is a large electron density in the intrinsic InAs layer with a peak electron density exceeding 10^{18} cm^{-3} (green dot-dash line). Moreover, the electron density is concentrated near the material interface, where the CB/VB overlap between the materials is maximised. The large carrier concentration in the InAs layer yields a high conductivity. Consequently, a memory device using bulk InAs as the channel layer will be NORMALLY-ON.

To realise a NORMALLY-OFF channel using this heterojunction, we consider the effect of reducing the thickness of the InAs layer such that a QW is formed. The resulting minimum occupation energy for the InAs CB will rely on its dimensions rather than the bulk bandgap of the material. The new minimum energy corresponds to the first energy Eigenvalue from the solving the Schrödinger equation for the well, where the bottom of the QW is the InAs CB minima. The thickness reduction increases the minimum energy necessary for an electron to occupy the QW channel (QW$_{\text{CH}}$). If the confinement energy is large enough, the energy overlap between the occupation energy for the InAs CB and maximum energy for the GaSb VB ceases. In other words, if the InAs thickness (d_{InAs}) is reduced dramatically (<8 nm), it becomes energetically impossible for electrons in the GaSb VB to occupy the InAs QW$_{\text{CH}}$ due to the confinement energy of the first level of the QW (E_{QW}).

Schrödinger-Poisson simulations (nextnano++) simulations are used to demonstrate this principle in detail [Fig. 4.9 **(b)-(d)**]. Here, an AlSb barrier is used to form the QW. Additionally, the GaSb layer in the simulation is p-doped to a concentration of 5×10^{16} cm^{-3}. In practice, GaSb possesses native accepting defects which occur regardless of growth conditions and technique. GaSb defects are formed from doubly accepting gallium antisites [221]. The concentration used here is similar to the doping level observed experimentally when forming GaSb via MBE under similar growth conditions [222].

The simulations demonstrate that a 10 nm InAs QW$_{\text{CH}}$ thickness provides insufficient confinement energy to produce the desired gap between the first QW$_{\text{CH}}$ energy state (E_{QW}) and the GaSb VB maxima [Fig. 4.9**(b)**]. Consequently, a transistor channel based on these dimensions

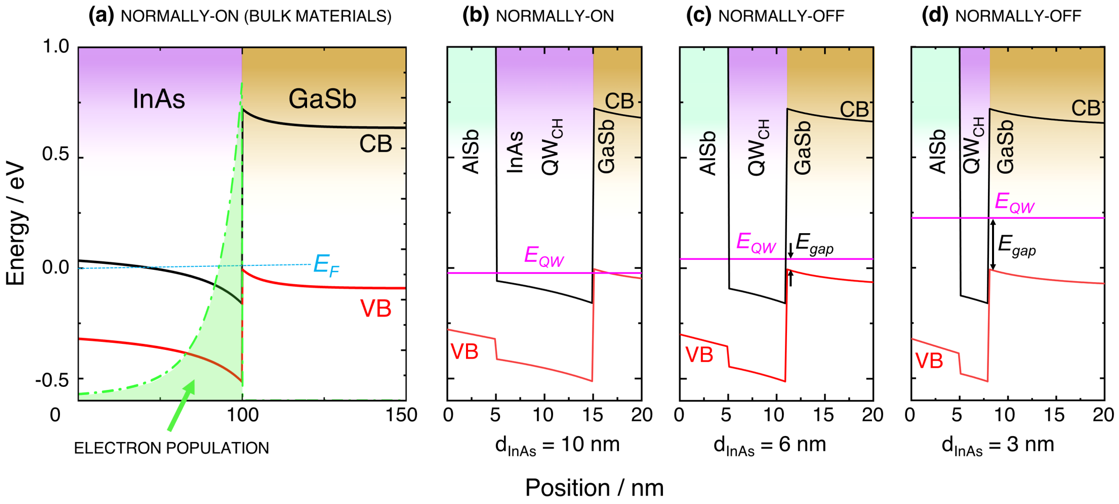


Figure 4.9: (a) Schrödinger-Poisson simulation of the bulk InAs/GaSb heterojunction (300 K, zero bias). The CB/VB overlap produces a large electron population in the InAs CB (green dot-dash line). **(b)-(d)** Schrödinger-Poisson simulations of QW channel (QW_{CH}) formation to produce an energy gap (E_{gap} between the GaSb VB and E_{QW} (pink lines). **(b)** 10 nm InAs (no gap). **(c)** 6 nm InAs. **(d)** 3 nm InAs.

would have a NORMALLY-ON channel. InAs channel thickness $<$ 8 nm raises E_{QW} above the GaSb VB to form an energy gap (E_{gap}), as shown in the results presented in Figures 4.9**(c)** and **(d)**. Electrons occupying the GaSb VB cannot traverse this gap; the InAs CB has effectively zero electron occupation. Consequently, with no carriers present, the InAs QW_{CH} is highly resistant under zero gate bias. As such, a transistor using this heterojunction with $d_{InAs} < 8$ nm will form a NORMALLY-OFF channel. Modulation of the channel conductivity will be discussed in detail in due course. However, the mechanism by which this is achieved is essentially a reinstatement of the InAs QW_{CH} / GaSb VB overlap by use of applied gate bias.

Detailed simulations of various channel dimensions were undertaken in which the future threshold voltages for channel current modulation were considered (subsection 4.5.2). This investigation concluded that the E_{gap} required for reliable MOS-like channel operation demands a d_{InAs} of 5 nm or less. We also consider introducing gallium into the channel to form a ternary alloy, $In_{1-x}Ga_xAs$. The introduction of Ga widens the bandgap [197], thus raising the energy of the channel CB minima with respect to the GaSb VB. The simulation results find that a small (20%, x = 0.2) concentration of Ga yields an E_{QW} - GaSb VB energy gap (E_{gap}) larger than that of the 5 nm InAs QW_{CH} at a 12 nm thickness. The inclusion of Ga into the channel allows for improved process tolerance but adds an extra growth parameter (composition), adding complexity to the memory fabrication including an increased lattice mismatch in the heterostructure.

4.6.1 nextnano++ simulation of the InGaAs/GaSb channel

In order to modulate the conductivity of the InGaAs/GaSb channel described in the previous section, we apply a positive voltage between the CG and BG of the device (Fig. 4.10). Doing so raises the VB maxima of the GaSb with respect to the $\mathrm{QW_{CH}}$. At a certain applied bias the GaSb VB energy matches the energy of the QW state in the channel (E_{QW}). From this point onwards, the InAs E_{QW} overlap with the GaSb VB is reinstated in a similar fashion to that described in the previous section. This allows electrons to suddenly pour into the InAs CB, thus increasing the conductivity of the channel layer. In brief, the voltage at which the GaSb VB maxima is equal to E_{QW} constitutes a threshold voltage (V_T) similar to that of MOSFET channels.

Schrödinger-Poisson simulations for the entire memory structure with a 500 nm GaSb layer are presented in Figure 4.10. Figure 4.10**(a)** shows the memory under zero bias from CG to BG (V_{CG-BG}). Here, there is a clear energy gap preventing electrons from occupying the $\mathrm{QW_{CH}}$. Figure 4.10**(b)** demonstrates the threshold voltage (V_T) at which the energy gap is eliminated. This occurs at around 0.03 V for this memory design. Figure 4.10**(c)** presents the band-structure with V_{CB-BG} bias increased beyond V_T. The increased energy overlap (ΔE) increases the number of electrons with sufficient energy to move into the InAs CB. Therefore, we predict that the electron population will continue to rise with V_{CG-BG} (and conductivity as a result).

After realising that the energy overlap (ΔE) can return by application of a gate voltage (V_{CG-BG}), the next step is to work out the corresponding channel conductivity for the gate-voltage sweep. Here, the question is: How many electrons shift from the GaSb VB to the InGaAs CB for each value of applied bias? Answering this question first involves simulating a V_{CG-BG} sweep to plot the ΔE dependence on V_{CG-BG}, which is carried out manually through individual measurements of each V_{CG-BG}. Once the relation between energy overlap, ΔE, and gate bias V_{CG-BG} has been obtained, we next seek to convert the calculated ΔE value into a meaningful channel conductivity relation by considering the DOS of this system.

The sides of the heterojunction have dissimilar dimensional properties due to their thicknesses. The InGaAs channel ($\mathrm{QW_{CH}}$) DOS is 2-dimensional; the density of states for electrons in the $\mathrm{QW_{CH}}$ is zero up until the first energy state of the QW is reached [$E_{QW}^{(1)}$, Fig. 4.11**(a)**] after which there is a sudden jump in available states which then steps upwards until at the next QW energy ($E_{QW}^{(2)}$, Fig. 4.11**(b)**), forming a step-like DOS described by Equation 4.2 of the previous section.

Schrödinger-Poisson calculations of the structure for a sweeping V_{CG-BG} are used to determine ΔE as a function of V_{CG-BG}. The results of this are presented in Figure 4.11**(b)**: the relation is linear with a threshold voltage (V_T) of ~ 0.03 V. As the applied voltage is increased, the GaSb VB energy (E_{GaSb}^{VB}) rises to overlap with high order QW states in the channel, $E_{QW}^{(2)}$ and $E_{QW}^{(3)}$ which are shown in Figure 4.11**(b)** by blue and green lines respectively.

The corresponding number density for the states is calculated directly from the previous

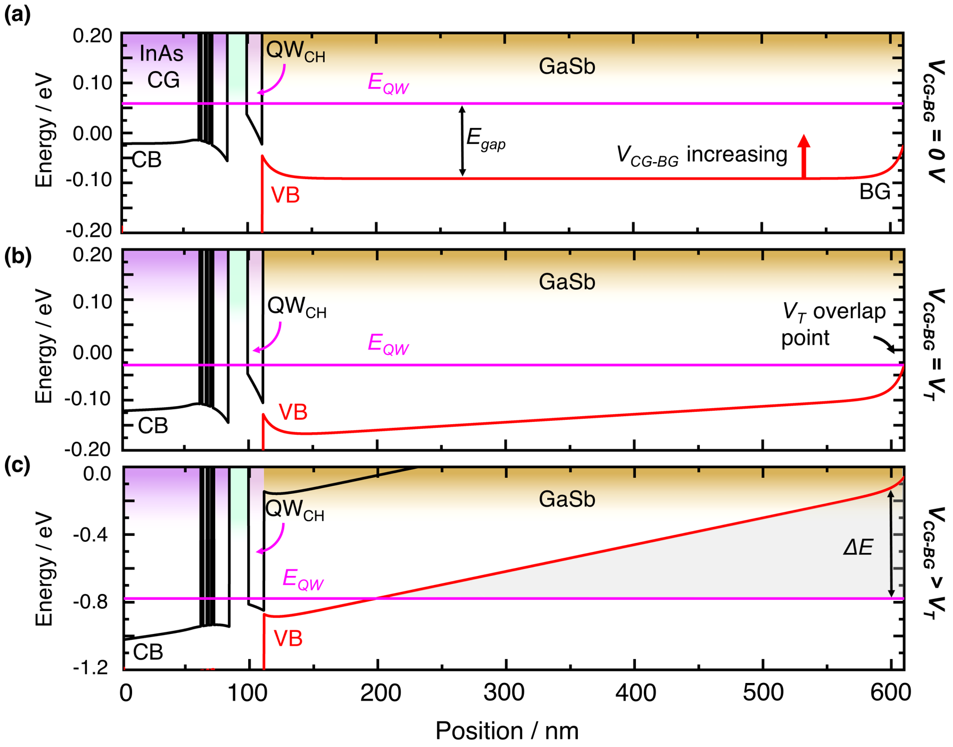


Figure 4.10: Schrödinger-Poisson calculations (nextnano++) for the memory structure (1D) demonstrating the channel conductivity modulation for the NORMALLY-OFF channel for CG-BG biases (V_{CG-BG}) of **(a)** 0 V, showing the energy gap between channel and GaSb VB. **(b)** V_T, the point where the energy gap ceases. **(c)** a larger voltage demonstrating the increased energy overlap with the QW$_{\text{CH}}$.

result using $m^*_{InGaAs} = 0.03052 m_e$ as the effective electron mass in the channel [224, 225] and is shown in Figure 4.11**(c)**. The position of the V_T is unchanged from the energy overlap as this is the origin of the modulation in number density due to the V_{CG-BG} at which the overlap begins. From here the beginning of the relationship is linear with increasing voltage (Equation 4.2) until the point at which overlap begins with higher order QW states ($E^{(2)}_{QW}$: blue shading and later $E^{(3)}_{QW}$: green shading). The additional energy overlaps contribute to the number density for the available states in the QW$_{\text{CH}}$ [Fig. 4.11**(b)**].

Following the calculation for the number density for a given applied V_{CG-BG}, it is straightforward to determine the conductivity relation of the 12 nm InGaAs channel for the above conditions. The electrons which fill the available states in the QW$_{\text{CH}}$ have a degeneracy factor of two. Thus, the electron density in the channel (n_e) is a function of applied gate bias from the number density calculated previously as

$$n_e(V_{CG-BG}) = 2n_{2D}(V_{CG-BG}) \,, \qquad (4.4)$$

where $n_{2D}(V_{CG-BG})$ is the density calculation presented in Figure 4.11**(c)**. Thereafter, this is trans-

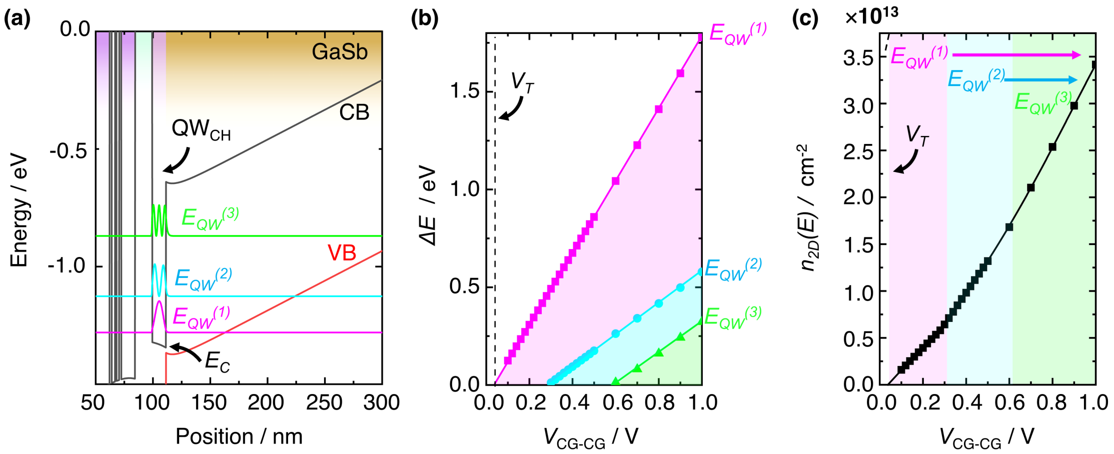


Figure 4.11: (a) Schrödinger-Poisson calculation (300 K) for the memory structure under V_{CG-BG} bias, demonstrating VB overlap with multiple QW energy states for the channel. **(b)** Energy overlap of the GaSb VB maxima with the energy states of the QW$_{CH}$ as a function of CB-BG bias, where each state: $E_{QW}^{(1)}$, $E_{QW}^{(2)}$ and $E_{QW}^{(3)}$ are represented separately as pink, blue and green lines respectively. **(c)** Number density for available QW states calculated from the energy overlap as a function of CG-BG bias (V_{CG-BG}). Coloured shading (pink, blue and green) correspond to the QW state that begins to add to the number density at that voltage.

formed into 2D conductivity using the Drude model [226], given as

$$\sigma_{2D} = e\mu n_{2D} \tag{4.5}$$

where σ_{2D} is conductivity, e is the charge of an electron and μ is the approximate mobility of electrons in the 12 nm InGaAs channel at 300 K taken from prior experimental literature [225]. Thus, the conductivity of QW$_{CH}$ is a function of bias applied from CG to BG (V_{CG-BG}) which is NORMALLY-OFF with a V_T around ~ 0.03 V (Fig. 4.12**(a)**, labelled logic 1 for empty FG).

Readout procedure

The modulation of channel conductivity with applied gate bias shares similarities with flash memory cells (subsection 2.2.1). This is not a coincidence as the device layers are engineered with memory readout in mind in order to appropriate the flash readout procedure. In the erased state (logic 1), the FG of the memory device is empty and the conductivity dependence on V_{CG-BG} is the result presented previously (Fig. 4.11). However, when the memory cell is programmed (*i.e.* the FG is filled with electrons) which is defined as logic state 0, the negative charges added to the FG screen the potential applied from CG to BG. Consequently, V_T shifts to a larger value; a larger positive voltage is required to produce the same electric field across the channel due to the presence of charges in the FG. The magnitude of the threshold voltage shift (ΔV_T) is given by

$$\Delta V_T = \frac{Q_{FG}}{C_{FG}} \, , \tag{4.6}$$

where C_{FG} is the capacitance between the FG and CG and Q_{FG} is the charge stored in the FG [54]. In this work, C_{FG} was calculated from a parallel plate approximation with the AlSb dielectric constant of ~ 14 [227] (C_{FG} = 0.8 µFcm^{-2}). The charge on the FG, Q_{FG}, is determined from simulation results which will be presented in due course. Note that as both Q_{FG} and C_{FG} are directly proportional to cross sectional area, it is eliminated from the above equation. This results in a one-dimensional equation for the threshold voltage shift, justifying the strictly 1D simulations used throughout the proceeding sections. Figure 4.12**(a)** presents the conductivity relation for the 0 (black line) and 1 (purple line) logic states of the memory cell, where to adjacent schematics demonstrate the corresponding FG conditions. The threshold shift for the program cycle is 430 mV, with a 0 state threshold voltage ($V_T^{(0)}$) of 0.46 V [$V_T^{(1)}$ remains at 0.03 V, Fig. 4.12**(a)**].

The threshold shift creates a system in which we now have different threshold voltages for the 1 and 0 logic states of the memory cell [labelled $V_T^{(1)}$ and $V_T^{(0)}$ respectively, Fig. 4.12**(a)**]. If we apply a reference voltage (V_{REF}) within the threshold voltage window[7], we can determine the logic state of the memory cell with an extremely high contrast [Fig. 4.12**(a)**]. The channel conductivity is significant if the memory cell is in state 1 and the conductivity is negligible if the cell is state 0 (although some leakage is expected in practice). Consequently, the logic of the memory cell can be read by measuring channel conductivity (current-sensing) whilst applying V_{REF} between CG and BG in a similar fashion to flash memory.

To demonstrate this procedure, Fig. 4.12**(b)** contains the simulation results for the bandstructure of the memory cell alongside the QW energies for the channel (with their probability density shifted for energy) under the condition; $V_{CG-BG} = V_{REF}$.. When the FG is empty (logic 1) the GaSb VB (solid red line) intersects the first QW energy ($E_{QW}^{(1)}$) to produce the energy overlap (ΔE) associated with an increased channel conductivity. On the other hand, the GaSb VB for a FG filled with electrons (logic 0) as shown by the dashed red line in the figure (energy shifted to align with the previous logic 1 VB calculation) lies significantly below the ground state of the QW$_{\text{CH}}$. Thus, there is no energy overlap resulting in a high-resistance channel layer.

Read disturb

During the readout, V_{REF} is applied across the CG-BG terminals of the device which will be combined with a small S-D voltage used to sense the conductivity of the channel for logic determination. As a result, an electric field is applied across the gate stack including the TBRT region. Due to the magnitude of V_{REF} ($\sim$ 0.3 V), combined with the increased distance over which the bias

[7] A voltage in between $V_T^{(1)}$ and $V_T^{(0)}$.

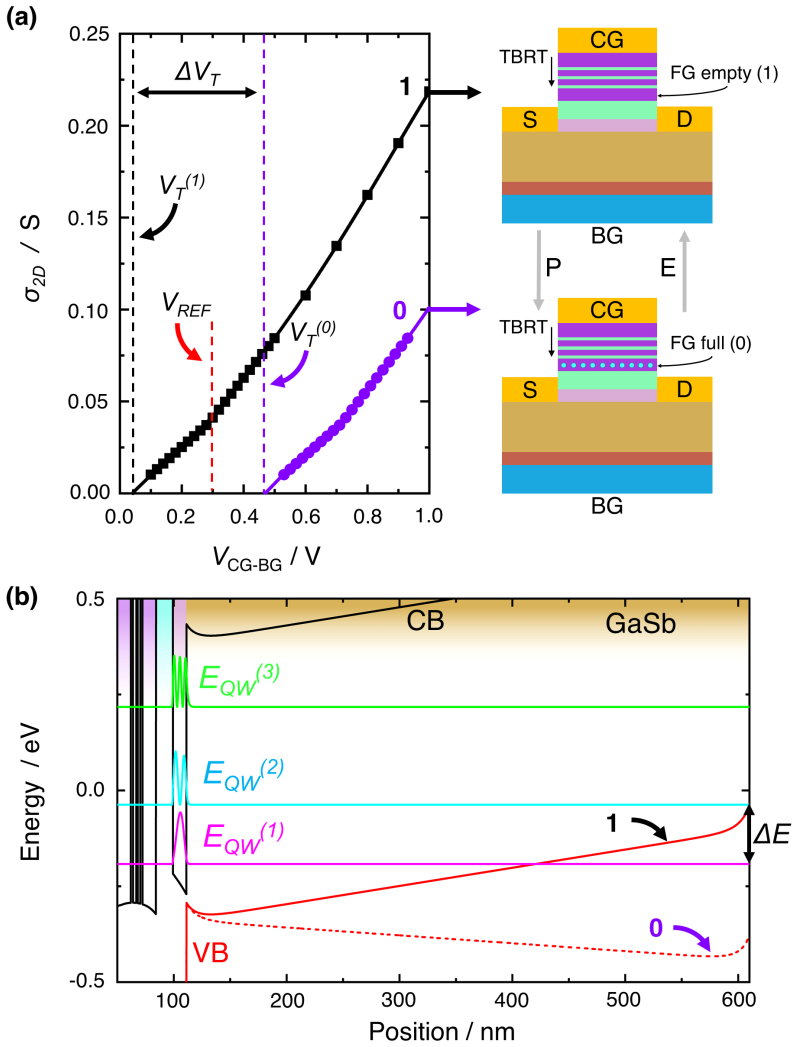


Figure 4.12: (a) Conductivity relation of QW_{CH} for an empty FG (logic 1) and a FG filled with electrons (logic 0). The reference voltage, V_{REF} dashed red line, is used to read the memory cell logic by channel conductivity measurement. **(b)** 300 K bandstructure calculations of the memory cell at $V_{CG-BG} = V_{REF}$ for logic state 1 (solid red VB line) and logic state 0 (red-dashed VB line, energy shifted to align with the 1 state calculation).

is applied, the resulting electric field across the tunnelling region of the gate stack is extremely small. In fact, the current-density simulations indicate that the readout voltage conditions produce a negligible disturbance to the FG logic ($> 10^{10}$ cycles before any meaningful disturbance occurs).

4.6.2 Combining program, erase and read using SPICE

A SPICE program (LTspice) was used to combine the P/E and read simulation results presented previously in order to determine the circuit level performance of the technology. There are many examples of SPICE models that have been used to characterise floating gate memories [62, 219, 228]. However, these are generally focused on modelling a device that has already been fabricated, extracting information for the model from experimental measurements such as capacitive coupling coefficients and tunnelling parameters (tunnelling parameters can also be modelled [62]). These are then inserted into the simulation to compare directly with experimental data [219].

In this work, where there are no established models or experimentally-derived parameters available, the data for the tunnelling mechanism is represented by a voltage-controlled current source (VCCS). The current (for a device area, A_{tun}) is modelled from a multiple-peaked asymmetric-Gaussian fit to the simulated tunnelling results. The result is dependent on the voltage applied across the tunnelling region. The voltage across the tunnelling region comes from two biases during the write and erase processes; the CG voltage and the source (S) voltage. The combined bias across the tunnelling region is determined from separate investigations of the band structure gradient (and resonant tunnelling alignments) using a Schrödinger-Poisson solver as detailed previously (4.5.2).

We next consider the voltage screening effect due to the presence of charge on the FG, which changes during the program or erase process. The current supplied by the VCCS changes as its own current output screens the input voltage, *i.e.* build up, or loss of, charge in the FG during write and erase pulses respectively. The simplest way to model this system is to connect the VCCS, containing all of the above information, to a capacitor with capacitance C_T, the total capacitance coupled to the FG from the tunnelling junction and charge blocking barrier (calculated from a parallel plate approximation as 2 μFcm^{-2}, Fig. 4.13). When a voltage pulse is applied, it is converted into the voltage across the tunnelling junction, from which the VCCS responds according to the resonant tunnelling simulation results of Figure 4.8**(e)** to release a current. This calculation is continuously cycled to take into account the changing charge on the FG during the voltage pulse. The electrons released in the program are stored on the FG (capacitor) and a voltage, V_{FG1}, is created (Fig. 4.13):

$$V_{FG1} = \frac{Q_{FG}}{C_T} \, . \tag{4.7}$$

This result then feeds back into the VCCS as a voltage screening effect. Similarly, this set up can be used to simulate charges leaving the FG (erase), where an initial voltage, $V_{INITIAL}$, defines the previously programmed state for the device. Combining equations 4.11 and 4.12 with the capacitances for the device, approximated as parallel plate capacitors using the layer thicknesses and dielectric constants of the materials, allows us to obtain an equation for the threshold voltage shift of the channel as a function of V_{FG1}, *i.e.*

$$\Delta V_T = \frac{C_T}{C_{FG}} V_{FG1} \tag{4.8}$$

The result is that we can track the threshold voltage shift for any given voltage pulse in a transient simulation to determine the change to the conductivity relation of the channel discussed in the

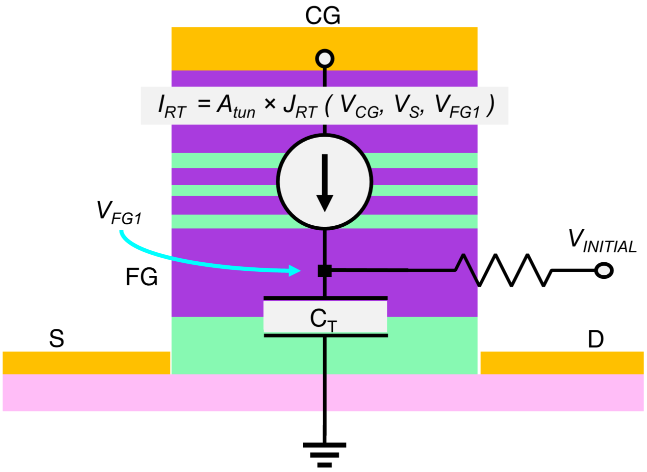


Figure 4.13: Schematic of the SPICE circuit-level simulation technique. Tunnelling current is given as a function of the CG voltage (V_{CG}), source voltage (V_S) and charge-screening voltage (V_{FG1}). $V_{INITIAL}$ allows us to add an initial screening voltage (used for the erase cycle).

previous section [Fig. 4.12((**a**)]. This allows us to investigate the performance of the memory and explore its circuit-level properties to form a suitable array architecture.

4.7 Simulation Results: Fast, Low-Energy NVRAM

4.7.1 Speed

The SPICE model results indicate that ULTRA**RAM**™ can operate at low voltage, low energy and high speeds. The transient simulations for the program cycle are shown in Fig. 4.14. Here, a voltage pulse (dark blue line) is applied to the CG (S-D grounded) and the response of the threshold voltage for the channel (black line) is calculated alongside the tunnelling current density for the resonant-tunnelling onto the FG (green line). The response of the model is highly dependent on the shape of the applied voltage pulse. Input pulses can be separated into three sections; the time taken to reach the target voltage (t_{rise}), the time the pulse remains at the target voltage (t_{on}) and the time taken to return to zero voltage (t_{fall}). For the purposes of this initial study these are each set to 5 ns (Fig. 4.21, blue line), *i.e.* $t = t_{\text{rise}} = t_{\text{on}} = t_{\text{fall}} = 5$ ns. For this pulse, with amplitude $V_{CG} =$ -2.0 V ($\sim$10 times smaller than Flash [75]), ΔV_T for the channel shifts by $\sim$0.8 V, which is sufficient to define the logic state 0 of the memory. Moreover, this process is mostly completed within the rise time (t_{RISE}) of the pulse, suggesting that faster switching is possible. The charge density, Q_{FG}, is the area under the tunnelling current density (j_{RT}) curve (green shading, Fig. 4.14) and is the sole reason for the change in ΔV_T in accordance with Equation 4.11.

It is important to note that within this model, capacitances C_{FG} and C_T are converted into

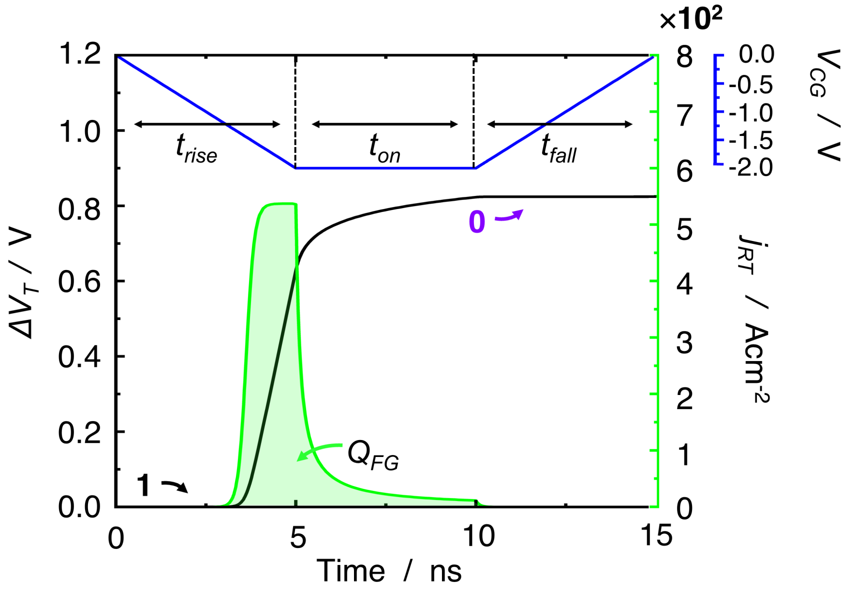


Figure 4.14: Transient SPICE simulations of the program cycle for an input CG voltage pulse of -2 V with a total time run time of 15 ns (blue). The resonant tunnelling current density into the FG during the pulse is shown by the green line which causes a threshold voltage shift (ΔV_T) in the channel (black line) due to the charge density stored in the FG (Q_{FG}).

areal capacitative densities to retain the strictly 1-dimensional properties of the model. With the model now established, we now seek to determine the maximum switching speed for the program cycle. Again, we define a voltage pulse in which rise time, on time and fall time are equal (t) and this time is varied to observe the simulation response. The transient simulations show that the channel threshold voltage can shift $\sim$ 0.35 V in under 100 ps for a t = 50 ps pulse [Fig 4.15**(a)**]. This switching speed is unprecedented for a non-volatile memory, which is attributed to the resonant tunnelling mechanism.

Interestingly, the current density results for the faster program cycles ($t < 1$ ns) have multiple peaks, whilst longer program cycles have a single peak [Fig 4.15**(b)**]. This is a result of the dual peaked TBRT current density relation [Fig. 4.8**(e)**] combined with the voltage screening effect. As electrons are added to the FG, their charge screens the applied voltage much like the CG-BG voltage (V_{CG-BG}) used for readout. Consequently, the voltage seen by the tunnelling region in the gate stack (V_{TBRT}) changes continuously during the cycle, where this decreases the voltage pulse for the program cycle (Fig. 4.16). Once the pulse is completed, the remaining screening voltage (V_{FG1}, labelled on Fig. 4.16) is, of course, the same screening voltage used to calculate ΔV_T (Equation 4.13).

For the erase cycle, an initial voltage ($V_{INITIAL}$) is placed on the FG to emulate a FG corresponding to a threshold shift of $\Delta V \sim 0.43$ V from a program cycle of 100 ps duration. The model is otherwise unchanged, with the erase current of the VCCS reversed to remove electrons from

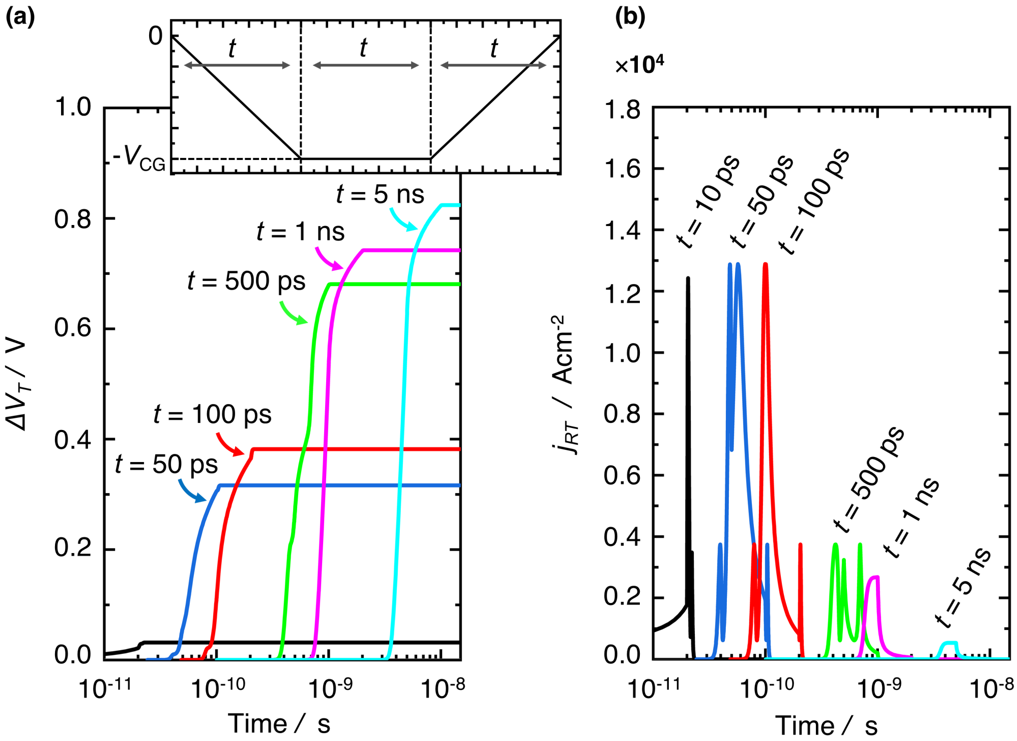


Figure 4.15: SPICE-model time dependence for the program cycle. **(a)** Threshold voltage shift for different voltage pulse durations, each with - 2 V magnitude on the CG. The graph inset defines the voltage pulse and the time value, t of each pulse, which range from 10 ps to 5 ns. **(b)** Corresponding current-density during the pulses of increasing t, which is the quantum transport into the FG during the cycle.

the FG. Transient modelling indicates that we can empty the FG (*i.e.* reverse the program cycle) within 200 ps for a 2.0 V CG pulse [Fig. 4.17**(a)**]. For longer pulses ($t > 500$ ps) the FG voltage becomes negative. However, this is an artefact of the simulation technique; current would become zero at $V_{FG1} \leq 0$ V as there are no electrons available to tunnel (FG empty). To clarify this, $V_T^{(1)}$ and $V_T^{(0)}$ are added to the graph [Fig. 4.17**(a)**].

The transient simulations for ULTRA**RAM**™ indicate that the 1-dimensional switching speed of the technology is $\sim$ 200 ps. However, one must consider the devices in 3-dimensions within a hypothetical array in order to make fair comparisons with current and emerging technologies. The 2.0 V P/E voltage is extremely low, and within the voltage available for a CMOS integrated circuit [229]. Thus, speed delays resulting from peripheral circuitry such as charge pumps (like in flash) are unlikely. The most important speed-limiting factor is capacitance, where the time taken to get to the target voltage (*i.e.* t_{RISE}) is limited by $\tau = RC$, which is the fundamental speed limitation of DRAM [75]. However, as a FG-memory, the capacitance of ULTRA**RAM**™ memory cells is extremely small (10^3 smaller than DRAM for a similar feature size). As a result, a t_{RISE} $\sim$ 200 ps is feasible for a 20 nm node, which is determined as follows: For a 20 nm DRAM cell,

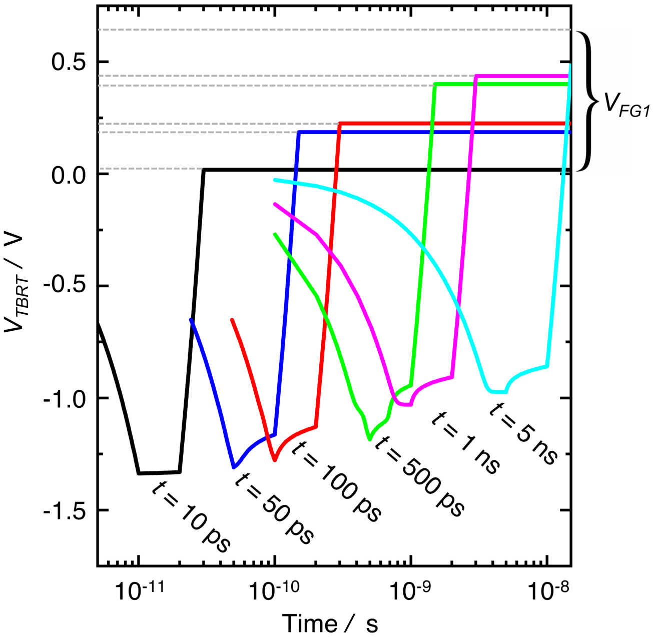


Figure 4.16: Simulation output for the voltage seen by the tunnelling region (V_{TBRT}) of the device over time from the program cycle. Voltage is reduced due to the screening effect of charges present on the FG. When the cycle is complete, the remaining voltage corresponds to the charges left on the FG (V_{FG1}).

$\tau_{DRAM} = RC_{DRAM} \sim 5$ ns , where $C_{DRAM} = 15$ fF. A similar feature size ULTRA**RAM**™ cell has capacitance $C_T = 8$ aF. Thus, assuming that the circuit resistance is of similar magnitude, the RC constant for ULTRA**RAM**™ will be at least three orders of magnitude less than DRAM; *i.e.* $\tau \sim 5$ps . However, it is possible that parasitic capacitance may provide limitations, should large arrays be implemented.

In Table 4.2, ULTRA**RAM**™'s switching performance is summarised alongside existing technologies. Although the speed of appears outrageous, it is simply a combination of a small-FG capacitance and low-voltage TBRT. As such, we can conclude that the performance is at least as good as DRAM, with the potential to approach SRAM speeds ($\sim$ 1 ns).

4.7.2 Energy

If we now compare some other important memory metrics for different types of memory cells with 20 nm feature size cell, we observe some striking results (Table 4.3). Most notable is the switching energy for the P/E cycling of ULTRA**RAM**™, which is lower than DRAM and NAND Flash by factors of 100 and 1000 respectively, and is significantly lower than other emerging technologies. This remarkable observation is a result of the combination of low voltages[8] and small capacitances.

[8] It should also be noted that the switching voltage of the simulations are corroborated by multiple experiments on large (10 µm) feature size devices [216].

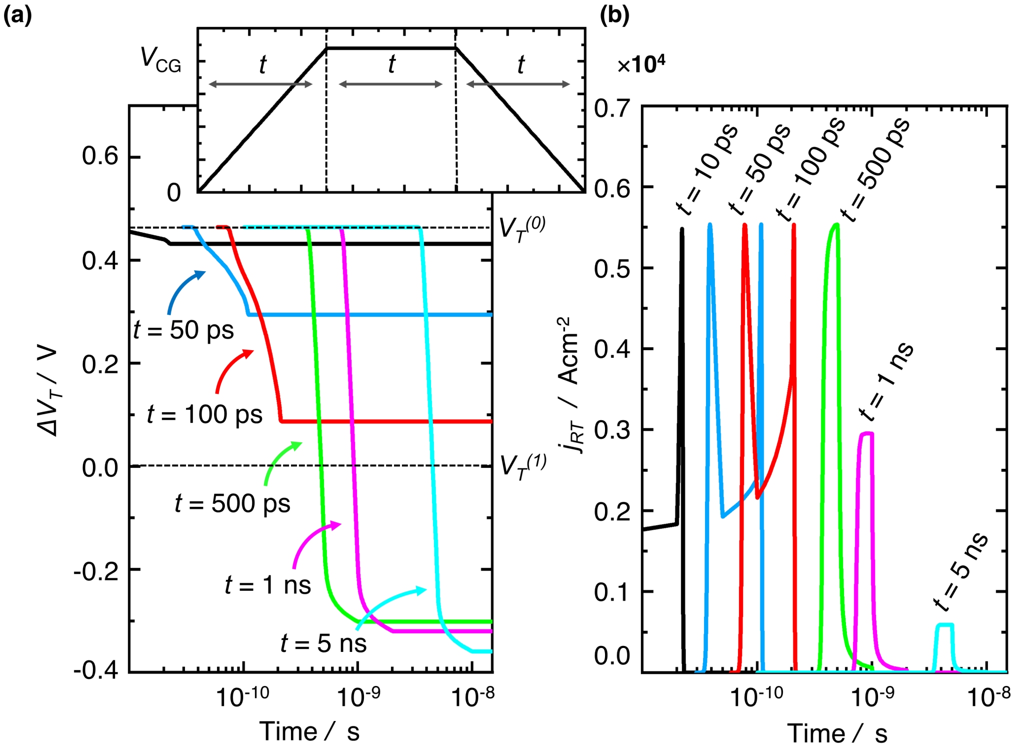


Figure 4.17: SPICE-model time dependence for the erase cycle. **(a)** Threshold voltage shift from a starting programmed state of $\Delta V_T = 0.43$ V for different voltage pulse durations, each with +2.2 V magnitude on the CG. The graph inset defines the voltage pulse and the time value, t of each pulse, which range from 10 ps to 5 ns. **(b)** Corresponding current-density during the pulses of increasing t, which is the quantum transport out of the FG during the cycle.

Furthermore, it contradicts the argument that non-volatility necessitates a greater expenditure of energy to change states than a volatile memory, due to the energy required to overcome barriers [1]. This is not the case for resonant tunnelling as the transport through the barriers occurs at very specific energy alignments, allowing us to have a high barrier energy (2.1 eV) but still observe tunnelling at small voltages.

The single concern within the benchmarking metrics listed in Table 4.3 is the electron number, which is the downside of the small FG capacitance. However, 2D NAND Flash technologies of similar feature size have just 30-50 electrons per cell level [230]. This comparison, combined with the high barrier energy and low disturb rate (discussed in detail in the proceeding subsections), suggests that this low number of stored electrons is not a stumbling block, at least not until the technology is scaled to feature sizes $<$ 10 nm.

Table 4.2: Switching speed limitations [75].

Technology	**Speed limitation**	**Switching time**
DRAM	Capacitor charging time	~ 10 ns
NAND	RC time constant to reach $\sim$ 25 V	$>$ 10 µs
PCM	Slow temperature ramp down to control crystallisation	100-400 ns
STTRAM	Stochastic switching due to spin precession	10-50 ns
ULTRARAM	**Time to charge FG via TBRT**	$<$ **1 ns**

Table 4.3: Benchmarking metrics for memories at the 20 nm node [75, 218].

Tech.	**Switching energy**	**Particle**	**Number**	**Barrier [eV]**
DRAM	$E = \frac{1}{2}CV^2$ $E = 0.5\times15\text{fF}\times0.6\text{V}^2$ $E \sim 10^{-15}$ J	Electron	$N = CV/q$ $15\text{fF}\times0.6$ V/q $\sim 5 \times 10^4$	0.55
3D NAND	$E = \frac{1}{2}CV^2$ $E = 0.5\times50\text{aF}\times20\text{V}^2$ $E \sim 10^{-14}$ J	Electron	$\sim 10^4$	1.6
PCM	$E = IVt$ $E = 0.1\text{mA}\times4\text{V}\times0.4\text{µs}$ $E \sim 10^{-10}$ J	Atomic bond Bond angle Bond coordination	$\sim 2 \times 10^4$	2.4
ReRAM	$E = IVt$ $E = 50\mu\text{A}\times3\text{V}\times50\text{ns}$ $E \sim 10^{-11}$ J	Cluster of oxygen vacancies or metal metal ions	10-1000	1.4-1.8
ULTRARAM	$E = \frac{1}{2}CV^2$ $E = 0.5\times8\text{aF}\times2.2\text{V}^2$ $\mathbf{E \sim 10^{-17}}$ **J**	**Electron**	$\sim$ **100**	**2.1**

4.7.3 Endurance

The endurance-limiting factor of FG memories such as flash are failures in the oxide tunnelling barrier due to the frequent application of high voltage where there are imperfections in the oxide [70]. The extremely low-voltage (and energy) used for P/E of ULTRA**RAM**™ should not produce the voltage-accelerated failures seen in flash, and the tunnelling region is grown as a single crystal with a small amount of strain such that imperfections are minimised. Moreover, RTDs using similar (InAs/AlSb) heterojunctions report no such device failures [231], despite operating at room temperature with oscillation frequencies up to 712 GHz and peak current densities of 2×10^5 Acm^{-2}: an order of magnitude greater than our tunnelling region [Fig 4.9**(e)**]. Thus, we expect the endurance of this memory to greatly exceed that of flash. However, rigorous testing on real-world devices is required to support this assertion, which is provided in later chapters.

4.7.4 Non-volatility

The ability of a memory to retain its logic state is important for memory performance in both RAM and mass storage applications [16]. It has been predicted that the intrinsic (300 K) storage time of electrons in the InAs/AlSb system exceeds the age of the universe [284]. However, this prediction is based on thermal excitation of electrons over the barrier potential. To investigate the non-volatility of ULTRA**RAM**™ we must consider the effects of the TBRT structure on the transparency of the barriers. This is accomplished by analysing the NEGF simulations of the TBRT region at 300 K under zero applied bias.

Fig. 4.18 shows the calculated CB edge (white) for the TBRT structure. The colour scale of the density of states (DOS) demonstrates the confinement energies of the QWs (QW_1 and QW_2). The transmission function, T, *i.e.* the likelihood of electrons leaving or entering the FG, is shown by the red line (log-scale). It is extremely small in the energy region below the barrier height; however, it possesses three distinct points of interest. The first is a transmission peak corresponding to the resonant state of the QW_1 ground state, the second is the transmission peak for the QW_2 ground state, and, finally, the largest transmission peak is for the second confined state of QW_1. Note that for our TBRT structure, QW_2 is too narrow to have a second confined QW state. The largest transmission peak of T = 0.04 resides at an energy (E) of around 1.8 eV, which corresponds to an electron storage time of about 10^{10} years at room temperature [284]. The lower energy (QW_1 and QW_2) transmission peaks are at $T \sim 10^{-5}$, E = 0.29 eV and at $T \sim 10^{-5}$, E = 0.4 eV, respectively. Although the peaks reside at a much lower energy, corresponding to millisecond storage times at room temperature for localization energies of that size [284], the probability of transmission is very low, making it unlikely that these peaks impact on the retention capability of the memory. Indeed, we will later observe fabricated devices which show stable memory retention exceeding 24 hours at 300 K with little or no state decay [233].

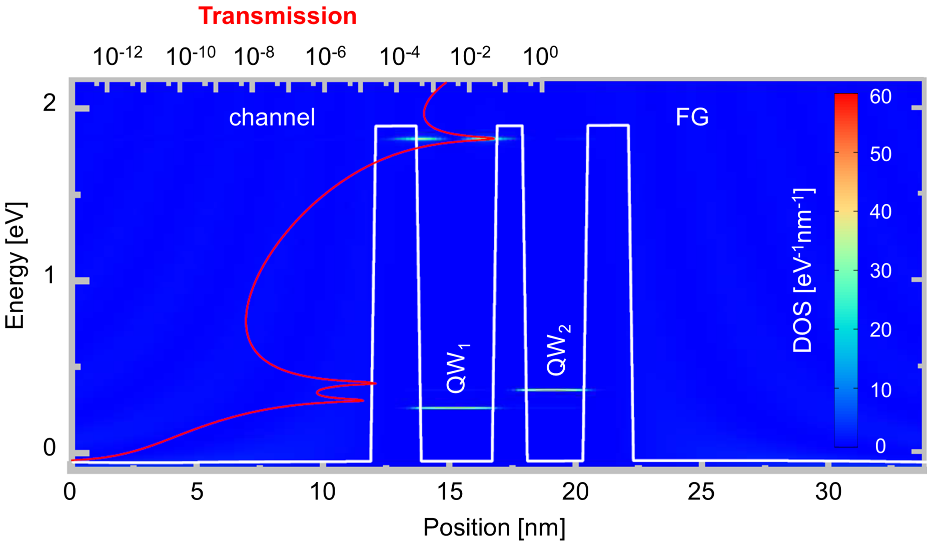


Figure 4.18: NEGF transmission calculations for the ULTRA**RAM**™ TBRT region at 300 K under zero bias overlaid on position-resolved, electron energy levels of the QWs for the target heterostructure, where the colour-scale indicates the DOS.The conduction band is shown by the white line. The corresponding transmission function (red line) demonstrates the peak alignments with the confined energy levels in the structure.

4.8 Proposed RAM Architecture

The similarities shared between ULTRA**RAM**™ and flash memories readily allows compatibility with flash architectures. One possible arrangement would be a NAND type architecture, with devices connected in series in large strings, producing a fast, low-power alternative to NAND-flash. However, large scale use would require 3D stacking to compete with the areal bit density of 3D NAND flash which is would not be straightforward for this technology. Instead, this arrangement could find a use in applications requiring smaller memory capacity, where reliable data retention, high speed and low energy is preferred to the high-bit density of NAND Flash such as in autonomous IoT sensors.

Alternatively, the devices could be implemented in a NOR-type architecture for use as an active memory (RAM). NOR-flash uses block-erase through CG voltage application, whilst single-cells are programmed via the HEI mechanism to add electrons to the FG (subsection 2.2.1).

The most important feature of an active memory is that it allows fast access to individual bits (devices) at the user's command [234]. Detailed investigations of the device operation from the SPICE-model allow us to realise a NOR architecture which permits the targetting of any cell for individual logic state change at any instance. The layout of this architecture is shown schematically in Fig. 4.19, where the common FGMOSFET symbol is replaced with the ULTRA**RAM**™ device symbol. The current-density peaks for the P/E process are very sharp, with extremely small current-densities below the alignment energies for resonant tunnelling. Thus, a significant voltage can be applied to a device terminal with negligible effect on the FG occupation. The array design exploits this property, in which half of the required voltage of a P/E cycle is applied to the target WL (CG) and the other half is applied to the target BL (S). These voltages combine on the target device to perform the P/E cycle. Of course, a half-voltage will be applied to every device which shares a common WL or BL with the target, however this does not compromise the data stored in these devices (disturb) for reasons previously outlined, and is discussed in more detail below.

Within the architecture (Fig. 4.19) the BG terminal serves as a common ground for all devices in the array (used for readout), with devices positioned back-to-back in pairs with grounded drain contacts. Consequently, the architecture requires only two interconnects (WL and BL) and one FG-device per bit with a shared drain; a very efficient and dense design capable of reaching a 4 F^2 bit area.

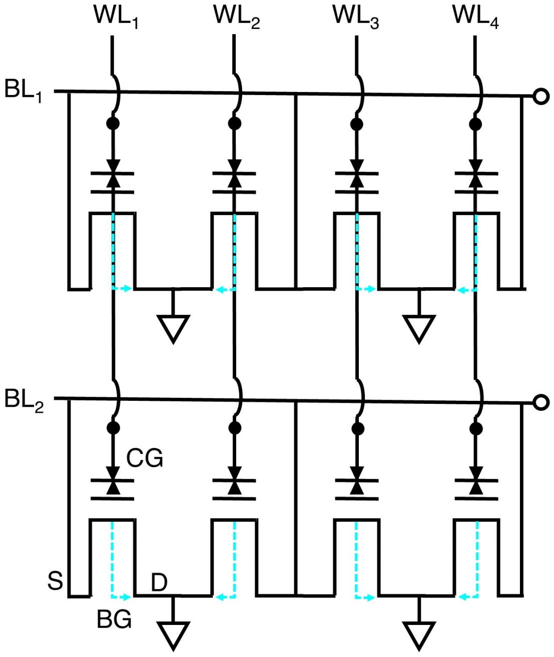


Figure 4.19: Circuit diagram for proposed ULTRA**RAM**™ architecture. Shared drain D and BG connections arranged in pairs optimises space on the chip. The ULTRA**RAM**™ circuit symbol is constructed from a combination of RTD and FGMOSFET circuit symbol.

Readout is very similar to flash memory, albeit with the use of the In(Ga)As/GaSb NORMALLY-OFF readout mechanism outlined in Section 4.6. A V_{CG-BG} voltage of V_{REF} is applied to the WL which will produce a conducting channel if logic 1 and high-resistance channel if logic 0. Simultaneously, a small voltage is applied to the target BL for channel-current sensing. The only device on the BL that could possibly be conducting is one which coincides with the WL voltage (V_{REF}), as all other devices are NORMALLY-OFF under zero gate bias. As such, the bit-line current measurement is isolated to measure only the S-D current of a device that has been targetted by selecting column and row (BL and WL). To summarise: if the measured BL current is high, the target device must be in logic 1, if a small current is measured, it is logic 0. Unlike the dominant RAM technology, DRAM, this readout procedure is non-destructive and will not disturb surrounding cells.

The half-voltage P/E procedure previously described is demonstrated within the SPICE model using $t = 1$ ns pulses. For the program cycle, a $V_{CG} = -1$ V (WL) and $V_S = 1$ V (BL) produces a ΔV_T of 0.63 V. The voltage applied to the source terminal (V_S) is corrected to the tunnelling voltage (V_{TBRT}) using the gate stack correction calculations similar to those detailed in subsection 4.4.2. This corresponds to a FG voltage (V_{FG1}) of -0.363 V, which is added to the FG prior to the erase simulation ($V_{INITIAL}$). Within the program simulation, we now apply the V_{CG} and V_S half-voltages separately many times, and observe how this effects the charge on the FG. After many half-cycles, the FG is partially programmed. However, the change in ΔV_T with respect to the number of cycles is an exponential decay, such that the disturb rate is reduced after a large number of half cycles [Fig. 4.20**(a)**]. This is a consequence of the voltage screening effect, whereby the FG voltage reduces the voltage seen by the TBRT region. Most importantly, the half-voltage disturb rate is such that at least 10^4 program cycles are required to partially shift the threshold voltage by $\sim$ 100 mV with $> 10^6$ half-cycles required to approach a logic state shift (*i.e.* more than half of the total ΔV_T is lost through disturbance). Note that this represents a disturb rate that exceeds the endurance of a flash cell and is an order of magnitude improvement over the DRAM disturb rate [235].

The half-voltage procedure is similar for the erase cycle [Fig. 4.20**(b)**]. Here we begin the simulation with the FG full (*i.e.* a potential is placed on the FG). Using voltage pulses for V_{CG} and V_S of 1.1 V and -1.1 V respectively, we observe that the FG is emptied using the half-voltage framework. We next apply the half-voltages individually in a similar fashion to the program cycle and observe if the logic state is partially erased by the lone half-cycles (disturb). As shown in Fig. 4.20**(b)**, 10^4 cycles produce a very small disturbance (15 mV) on the logic state. Moreover, there is almost a negligible disturbance from the half-voltage on the source terminal (V_S). The relation of the disturbance is again an exponential decay, for reasons similar to those described for the program cycle. Consequently, it is predicted that the memory cells can easily withstand 10^6 half-voltage erase cycles without many disruption the logic 0 state.

In summary, the half-voltage architecture is shown to work with program voltages on the CG and S terminals of -1 V and 1 V respectively and erase voltages on CG and S of 1.1 V and -1.1 V respectively. Short $t = 1$ ns pulses produce sufficient threshold shift to define logic states 0 and 1. The corresponding disturb rate within the RAM architecture is small enough to preserve the given logic for up to 10^6 P/E cycles on shared WLs or BLs.

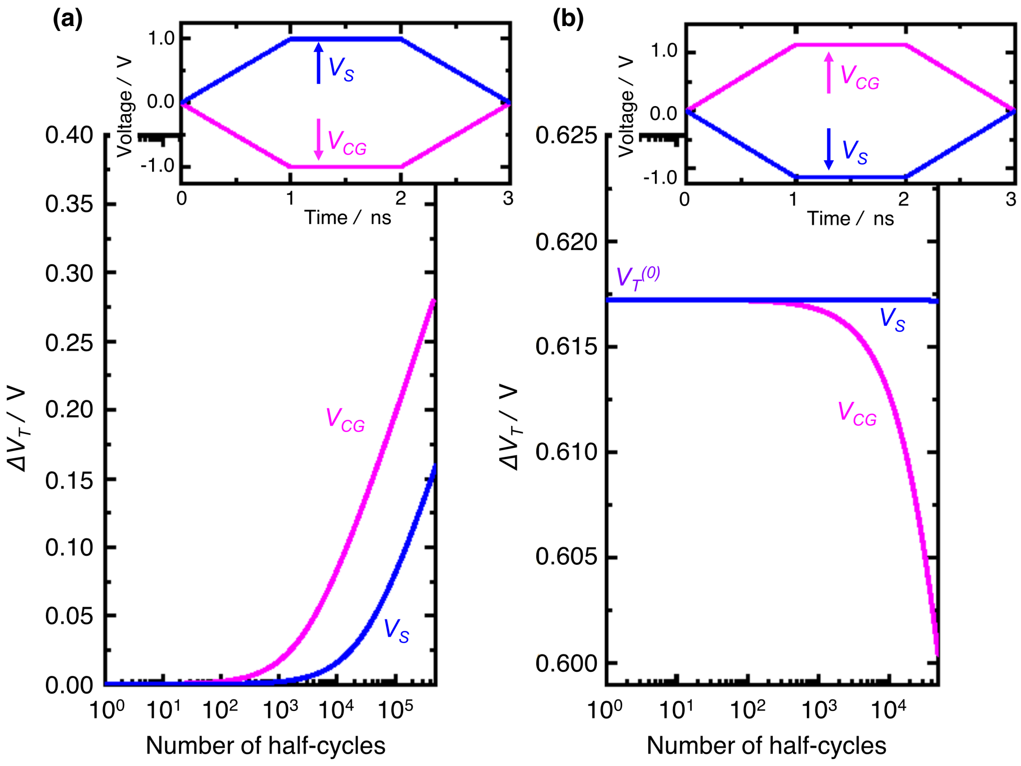


Figure 4.20: Single half voltage simulation results for threshold voltage shift from **(a)** programmed state **(b)** and erased state **(b)**. Half-voltages are applied to CG (V_{CG}) and S (V_S) and are shown in the inset whereby a lone $t = 1$ ns programming half-voltage is applied to the cell many times (*i.e.* disturb for the architecture).

4.9 Elimination of gate leakage current

Experimentally, initial prototype single cell devices [216] have exhibited a limited endurance despite operating at low voltages (< 2.5 V). These devices also demonstrated a large current through the entire structure (CG-BG and CG-S/D) despite the 2.1 eV CB offset of the InAs/AlSb heterojunction. Although the InAs/AlSb system provides an excellent CB offset, the VB offset is just 0.1 eV [197]. This undoubtedly allows holes to travel freely through the device, creating large currents from CG to BG during P/E cycling which cause gradual device degradation. Unfortunately, a material with a large VB offset to InAs does not exist within the 6.1-Å semiconductor family. Instead, we must amend the design of the devices such that the position of the TBRT region is reversed, as shown schematically in Fig. 4.21. This allows us to opt for a gate-last processing technique in which a charge blocking barrier is fabricated from a layer of ALD-deposited Al_2O_3, providing the necessary band-offsets with InAs to block both electrons and holes (CB 3.1 eV, VB 2.1 eV [205]). With this problem eliminated, it is predicted that the endurance of the new iterations will vastly improve.

The alteration brings about some obvious changes to the simulations previously presented.

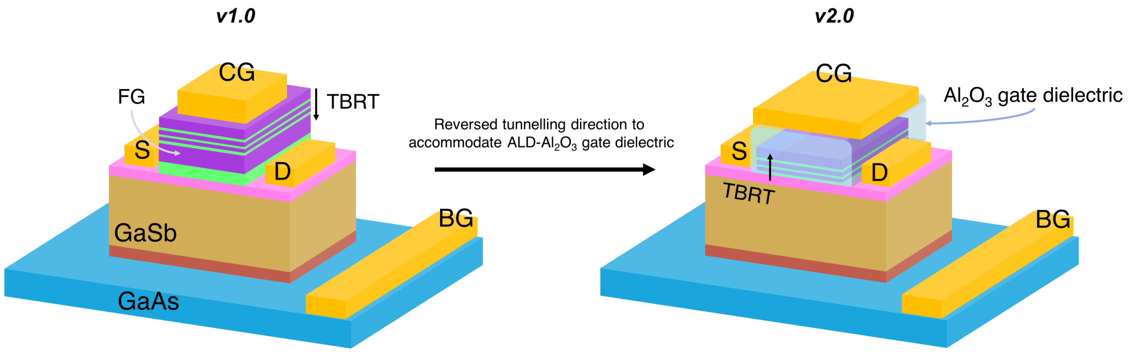

v1.0
CG
FG
TBRT
S
D
GaSb
GaAs
BG
Reversed tunnelling direction to accommodate ALD-Al2O3 gate dielectric
v2.0
CG
Al2O3 gate dielectric
S
D
TBRT
BG

Figure 4.21: Device design of the simulations results presented previously (*v1.0*, left) followed by the new design with the TBRT region reversed in order to accommodate the Al_2O_3 charge-blocking barrier (*v2.0*, right).

Firstly, the capacitances of the devices will be adjusted, as will gate stack corrections for the applied voltages. However, the ALD-layer is high-k ($\sim$ 8 [236]) and its thickness can be selected with atomic-layer precision such that the overall capacitance and necessary voltages are not significantly altered from the previous calculations. Secondly, polarities of applied voltages are reversed for P/E cycling, as tunnelling into the FG occurs in the opposite direction. Lastly, one must be sure that there is sufficient population of electrons in the channel during the program cycle, as the NORMALLY-OFF channel layer is now the source from which electrons must originate to undergo resonant tunnelling. Fortunately, the program cycle requires a positive bias on the CG for the program to take place in this design. Consequently, the grounded BG will allow a sufficient V_{CG-BG} to populate the channel, thus sourcing electrons for tunnelling into the FG. In other words, the electrons programmed into the FG originate from the GaSb VB, relying on an ON channel configuration during the program cycle.

4.10 Channel modelling revisited

Initial testing of devices with a 500 nm thick GaSb layer (*v2.0*) gave extremely high resistance channel measurements (10^{11} Ω) at all CG-BG biases. The possibility that this was due to experimental factors such as growth or fabrication errors was systemically eliminated by cross-sectional TEM of the growth and post-process cross-sectional imaging of the devices (BEXP). Consequently, we conclude that the electrons in the GaSb VB must be unable to access the In(Ga)As for an unforeseen reason, which prompted a more detailed revisiting of the channel modelling.

When formulating the concept of the NORMALLY-OFF channel, the calculations are predicated on one fundamental assumption: If electrons in the GaSb VB have sufficient energy, they will travel into the $In_{0.8}Ga_{0.2}As$ CB (QW_{CH}). By this reasoning, we assume that the raising of the GaSb

VB energy relative to the QW_{CH} raises the Fermi level of the system and electrons move into the channel by overcoming the effective bandgap energy created by the quantisation (similar to that described in [204]). If this is the case, the channel conductivity dependence will be independent of GaSb layer thickness (d) as previously calculated for the 500 nm layer. However, quasi-Fermi level[9] calculations indicate that the Fermi-level (Fig. 4.22**(a)**, cyan line) in the immediate vicinity of the QW_{CH} remains below its quantised energy state (pink line). This indicates that our previous assumption is incorrect and the channel conductivity modulation must therefore possess a dependence on the GaSb layer thickness [d, Fig 4.22**(a)**].

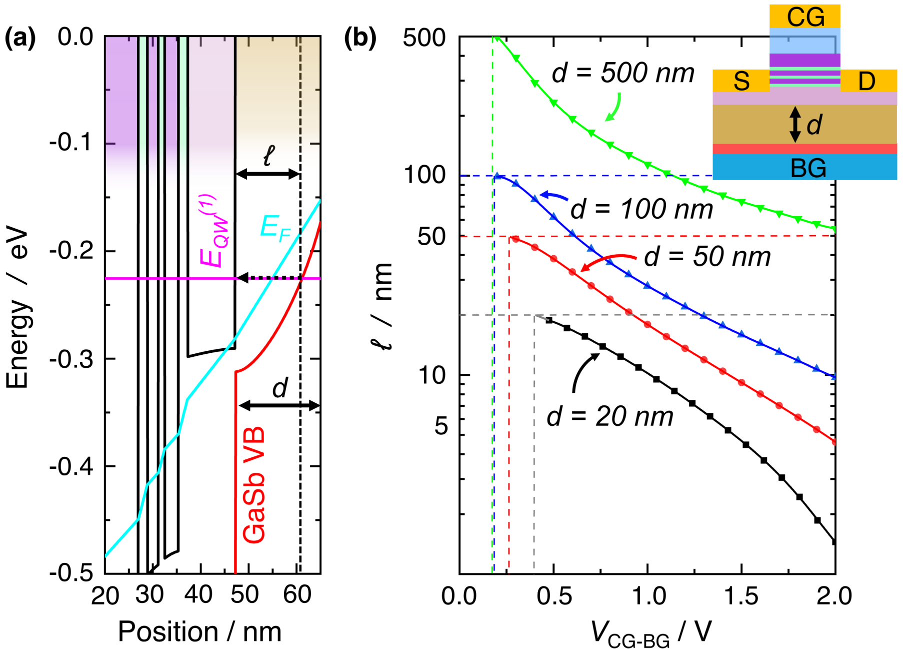


Figure 4.22: Revisited channel modelling: **(a)** Memory bandstructure simulation (300 K) with V_{CG-BG} applied, including CB (black line), GaSb VB (red) quasi-Fermi level calculation (blue line) and QW_{CH} first quantised energy level. The tunnelling distance (ℓ) is defined as the distance from GaSb VB at the energy of the channel layer QW_{CH} state. **(b)** Tunnelling distance (ℓ) dependence on CG-BG voltage for readout for multiple GaSb layer thicknesses, d, as shown in the inset schematic.

Given the position of the quasi-Fermi level, the mechanism of electron transfer into the channel layer from the GaSb VB must proceed by means of band-to-band tunnelling, akin to tunnelling FETs (TFETs) using similar material systems and thicknesses [238]. The electron tunnelling probability relies on a short tunnelling distance, in which the likelihood of tunnelling decays exponentially as a function of separation [237]. In our case, this is the length spanning from the QW_{CH} to the GaSb VB at the minimum required energy ($E_{QW}^{(1)}$). This distance is shown in the Schrödinger-Poisson calculation presented in Figure 4.22**(a)**, labelled ℓ, noting that the tunnelling distance is to

[9] The Fermi level is defined for calculations in equilibrium - applying a voltage across the structure produces a non-equilibrium (or 'quasi') Fermi level.

the GaSb VB, as the quasi-Fermi level lies in bandgap where there are no carriers available.

Figure 4.22**(b)** plots the tunnelling distance, as previously defined (ℓ), as a function of CG-BG potential (V_{CG-BG}) for various choices of GaSb layer thickness (d) ranging from 500 nm to 20 nm. Each data-point for ℓ is collected from the Schrödinger-Poisson calculations for the given d and V_{CG-BG} values. The results show that the initial design ($d = 500$ nm) has a tunnelling distance $>$ 50 nm at 2 V. Consequently, the tunnelling probability for electrons moving into the channel will be negligible. This is consistent with the experimental findings, as the channel remains NORMALLY-OFF. For the thinnest GaSb layer ($d = 20$ nm) the tunnelling distance is just a few nanometers, which greatly increases tunnelling probability and therefore provides conductivity modulation in the channel. This is consistent with previous experiments on TFETs using this heterojunction with similar dimensions [178, 204, 239]. Accordingly, a 20 nm GaSb layer was used in following iterations of the technology (*v2.1* onwards). It is important to note that this design is not a TFET, where the band-to-band tunnelling current *is* the S-D current. Here, the band-to-band tunnelling modulates carrier occupation of the channel by use of gate voltage, and the S-D conductivity reacts to the change in electron occupation.

4.11 Summary

In this chapter, a III-V semiconductor NVRAM with startlingly low switching energy (10^{-17} J) for a 20-nm feature size) that operates as an FG memory at significantly lower voltages than Flash ($\leq$2.3 V). Positive endurance and data retention results are expected due to the extremely low switching energy and large barrier energy (2.1 eV), although testing of this on experimental devices is required. The combination of nextnano.MSB, nextnano++, and SPICE simulations indicates that the device can operate virtually disturb-free at sub-ns pulse durations, at least an order of magnitude faster than the volatile alternative, DRAM. These advantages are derived from the triple-barrier RT mechanism used to transport the charge in and out of the device, which occurs at much lower voltages than other FG memories (*i.e.* flash). The proposed device has a threshold voltage and threshold voltage shift due to charge storage, allowing a similar read process to that of FGMOSFET cells used in Flash memory. This is achieved using a broken gap (Type-III) conduction band alignment formed from an $In_{1-x}Ga_xAs$/GaSb heterojunction, where the $In_{1-x}Ga_xAs$ channel is a thin (12 nm) QW. An excellent contrast in threshold voltages between the 0 state and 1 state is achieved. The resemblance to flash memory cells allows NAND or NOR Flash architectures to be directly implemented on the device to produce large arrays. The simulation results indicate that half-voltages can be used within a NOR-type architecture to target individual cells for write, erase, and read processes. This exclusive feature, combined with the increased speed suggested from the transient results of the 1-D model, predicts that the device can be implemented in large arrays as a low-power, non-volatile, non-destructively read alternative to DRAM.

Chapter 5

Memory Fabrication

In the previous chapter, the operation and design of ULTRA**RAM**™ memories is described in detail. Next, we endeavour to fabricate this design with the goal to demonstrate memory operation with electrical measurements on microscopic single devices and small (2×2) memory arrays.

5.1 Growth Details

In this work, the ULTRA**RAM**™ memory heterostructures are grown on 2-inch highly n-doped (Si doped, $n \sim 2 \times 10^{18}$ cm^{-3}) GaAs and 3-inch n-doped Si wafers (phosphorus doped, $n \sim 2 \times 10^{18}$ cm^{-3} with 4° offcut). The MBE layers grown on these substrates have substantial lattice mismatches of ~ 7 % and ~ 12 % for GaAs and Si respectively. The lattice disparity is resolved by use of an interfacial misfit array (IMF) between the substrate and GaSb buffer layer. The IMF layer results in a lattice-mismatched interface whereby strain energy caused by lattice mismatch is alleviated by misfit dislocations which propagate laterally, *i.e.* dislocations are confined to the epilayer/substrate interface [240]. The cross sectional TEM imaging of a GaSb/GaAs interface using an GaSb IMF layer is shown in Figure 5.1, and convincingly demonstrates this principle. The dislocations propagate laterally to form a zip-like pattern on the cross-section. Crucially, the material above the buffer remains high-quality, atop of which the ULTRA**RAM**™ heterostructures are formed.

For GaAs substrates, an IMF GaSb layer grown at optimised growth conditions (which include Sb flux and substrate temperature) greatly reduces defect density within subsequent material layers of the 6.1-Å semiconductors [241, 242]. Since the initial prototype devices (*v1.0*), the technique has been further optimised and subsequent materials appear to have significantly improved quality, including the complete removal of large oval defects which were numerous beforehand. The details of the optimisation are summarised in Table 5.1. All other growth parameters for GaAs substrates are otherwise unchanged from the initial prototypes, albeit with different target

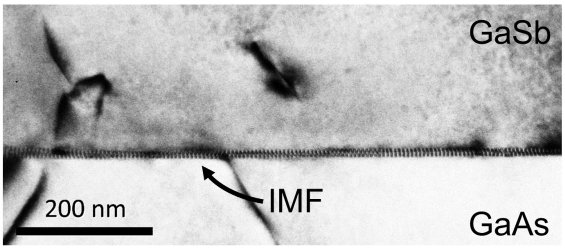


Figure 5.1: Cross-sectional TEM image of the GaSb/GaAs interface using a GaSb IMF layer to laterally confine misfit dislocations. This is a GaSb/GaAs/GaSb/Ge/Si test sample, hence the threading dislocations present in the GaAs layer propagate from the GaAs/Ge interface. Image provided by Prof. Richard Beanland (University of Warwick, UK) with permission.

layer thicknesses for the memory structure. Thus, we will not repeat the details of this procedure here, which are available in [216].

The growth of GaSb on Si is a technique in which AlSb islands are used to grow GaSb directly on Si wafers (via MBE) [244]. Here, the GaSb buffer layer was first grown on a Si (100) substrate with a 4° offcut toward (011) using a thin 17 monolayer (ML) AlSb nucleation layer. The quality of bulk GaSb grown on Si is enhanced by the formation of three-dimensional (3-D) AlSb islands on the Si surface, which reduces the diffusion length of Ga atoms, thus promoting planar growth of GaSb. Moreover, the strain at the Si/GaSb interface is relieved by formation of an IMF layer, to the same effect of the GaAs/GaSb interface as detailed previously. This technique significantly reduces surface defect density and surface roughness compared to a trial where GaSb was grown on a Si/Ge wafer (supplied by IQE) using a GaAs buffer on the Ge[1] before implementing the GaAs/GaSb growth technique. Growing GaSb directly on Si also eliminated the large oval defects on the surface (Table 5.1). First-attempt growths had their defect density[2] reduced even compared to the growths on GaAs substrates, albeit with an increased surface roughness. However, the Si/Ge/GaAs/GaSb growth procedure was not fully calibrated or optimised, so it is not possible to say which growth technique is superior for implementing ULTRA**RAM**™ on Si from this result.

Analysis of the TEM from initial memory growth on GaSb/Si showed that many threading dislocations propagate through the GaSb buffer and most terminate within the first 500 nm. However, many are terminated at the InAs/AlSb interface at the BG of the memory design [Fig. 5.2**(a)**], leaving high quality layer growth above for memory operation. In light of this, subsequent growths on Si implemented extra GaSb/AlSb heterostructures below the BG to act as dislocation filters (DFs). This is similar to the work of Delli *et al*, where many InAs/AlSb layers were grown prior to epitaxy of InAs/InAsSb quantum wells for mid-infrared light sources [245], significantly improv-

[1] GaAs is lattice matched to Ge.

[2] Measured by electron channelling contrast imaging (ECCI).

ing material quality. Here, for the growth used for ULTRA**RAM**™ *v2.3*, four DFs were used each separated by 100 nm of GaSb, consisting of five repeats of GaSb (10nm) and AlSb (10nm) [Fig. 5.2**(b)**]. The inclusion of dislocation filters was found to reduce the number of stacking faults and reduce the defect density (Table 5.1). However, they are still occasionally present in the memory layers. The improvement in material quality through the introduction of DFs provides a step in the right direction, with a clear path towards improving material quality for ULTRA**RAM**™ on Si. Many dislocations appear in the upper memory layers, despite the underlying material having better quality (Fig. 5.2, bottom images): The relatively thick (>30 nm) InAs layer used for the BG is likely to have exceeded the critical thickness, generating misfit dislocations through the memory layers above it. This issue could be resolved by thinning the InAs BG layer, which would require improved process control, or substituting it for a lattice-matched material with free carriers at similar energy (p-GaSb). The growth for *v2.3* also includes a 1.2 nm layer of AlSb between the InAs/GaSb interface used for channel operation. This layer acts as an etch stop for the selective wet etch, which will be discussed in detail later.

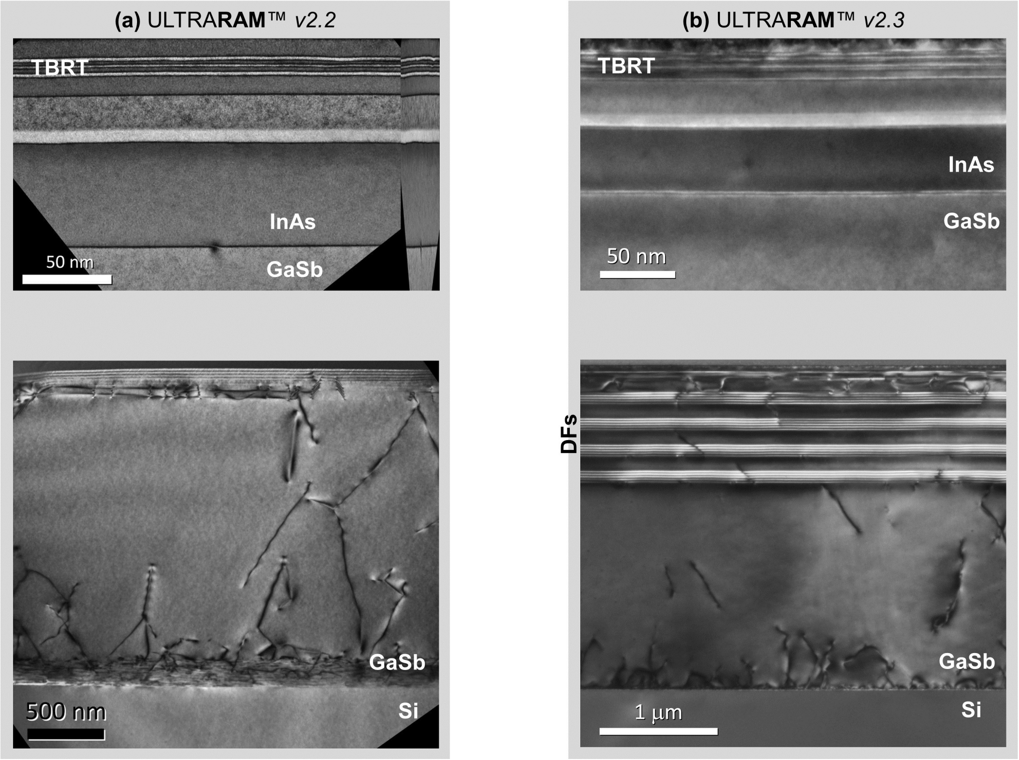


Figure 5.2: Cross-sectional TEM imaging for comparison of memory growths of Si for **(a)** ULTRA**RAM**™ on Si *v2.2* and **(b)** ULTRA**RAM**™ on Si *v2.3* with DFs and an AlSb layer in the channel. Images provided by Dr. Richard Beanland (University of Warwick, UK) with permission.

Table 5.1: MBE growth details in chronological order demonstrating material improvements of GaAs and Si substrates. Defect density is measured by electron channelling contrast imaging (ECCI) and was carried out by Prof. Richard Beanland (University of Warwick).

	IMF conditions				
Growth	**Substrate temp (°C)**	**Sb flux (ML/s)**	**AFM roughness (nm)**	**Optical microscope**	**Surface defect density (cm^{-2})**
Initial memory on GaAs *v1.0* [216]	500	2.7	-	oval defects 10^5cm^{-2}	-
GaSb/GaAs test	505	1.8	0.86	none visible	-
ULTRA**RAM**™ on GaAs *v2.1*	505	1.8	~ 0.80	none visible	5×10^8
GaSb/GaAs/Ge/Si trial	505	2.0	3.5	oval defects 10^4cm^{-2}	7.8×10^9
InGaAs/GaSb/AlSb/Si	485*	2.3	1.7	none visible	2×10^8
Memory on GaSb/AlSb/Si	485*	2.3	2.3	none visible	-
GaSb/AlSb/Si	485*	2.3	1.3	none visible	2×10^8
ULTRA**RAM**™ on Si *v2.2*	485*	2.3	X	none visible	2.5×10^8
ULTRA**RAM**™ on Si with DFs *v2.3*	485*	2.3	X	none visible	1.4×10^8

* indicates two-step buffer growth described in [244].

5.1.1 GaAs Substrate *v2.1*

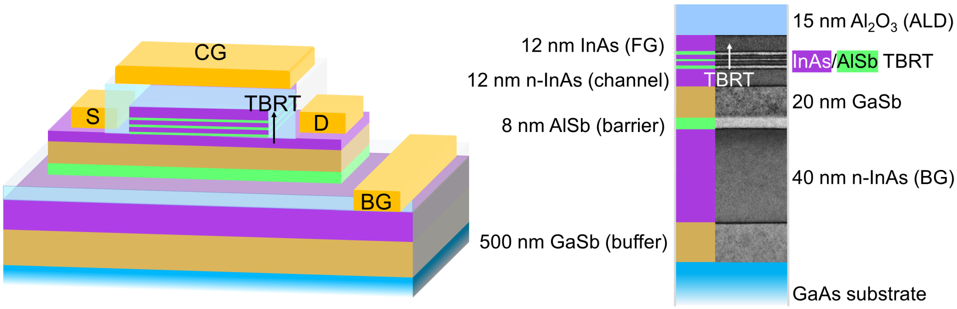


Figure 5.3: Schematic of the *v2.1* ULTRA**RAM**™ design with corresponding layer thicknesses. The Al_2O_3 gate dielectric layer is added later via ALD (*i.e.* MBE growth is terminated at the 12 nm InAs layer which defines the FG). TEM cross-sectional images of the MBE-growth are presented alongside.

The design for ULTRA**RAM**™ *v2.1* is shown in schematically in Figure 5.3 alongside cross-sectional TEM characterisation of the wafer; clearly demonstrating the ULTRA**RAM**™ heterostructure. Generally, the GaSb layer grown on the mismatched substrates must be thick to ensure material quality. Dislocations frequently decay in the buffer layer to leave a high quality surface. Consequently, thin (20 nm) GaSb layers for channel operation as described in the previous chapter (*v2.1* onwards) cannot act as the GaSb buffer layer and must be grown separately. As such, these growths include a buried n-doped InAs layer for the back-gate above the initial GaSb buffer. Then, an AlSb layer of 8 nm thickness is grown to act as a charge blocking layer separating the BG from the channel. The 20 nm GaSb layer is then grown on top on top of this, followed by the rest of the memory heterostructure (Fig. 5.3). This design necessitates front-side BG contact, as the BG layer is buried within the material layers.

The channel layer for *v2.1* is doped (n-type). This will cause the channel to be NORMALLY-ON, reducing the logic readout contrast. However, this also simplifies the device operation, allowing the basic memory functions to be analysed independently from the readout mechanism. Importantly, the processing and growth quality can be assessed before implementing more complex channel designs.

5.1.2 Si Substrate

Growth of memory heterostructures on Si substrates have similar layers near to the surface, where the important resonant tunnelling memory operation occurs. As demonstrated in Fig. 5.4, the key difference is the layers used to implement high quality BG, channel and TBRT layers on the Si substrate. ULTRA**RAM**™ *v2.2* [Fig. 5.4**(a)**] features a two step GaSb buffer growth as the foundation for epitaxy of the memory structure, whereas *v2.3* [Fig. 5.4**(b)**] adds multiple GaSb/AlSb superlattices to act as dislocation filters. The channel of *v2.2* is n-doped InAs, which is again a NORMALLY-ON design to assess basic memory operation on Si substrates. The channel of *v2.3* is 5 nm InAs to form the QW channel described in the previous chapter. Note that a sample with a $In_{0.8}Ga_{0.2}As$ channel (as described in the previous chapter) was attempted. However, the increased lattice mismatch resulted in a higher defect density in the channel which lead to processing difficulties. A thin (1.2 nm) AlSb barrier is added between the InAs QW and the GaSb layer. This improves process uniformity, as it acts as an etch stop to protect the GaSb under the channel when wet etching, which occurs due to etch-pitting of the InAs layer. Furthermore, Poisson-Schrödinger simulations indicate that the inclusion of the barrier has minimal impact on the position of the channel QW ground state (<1 meV).

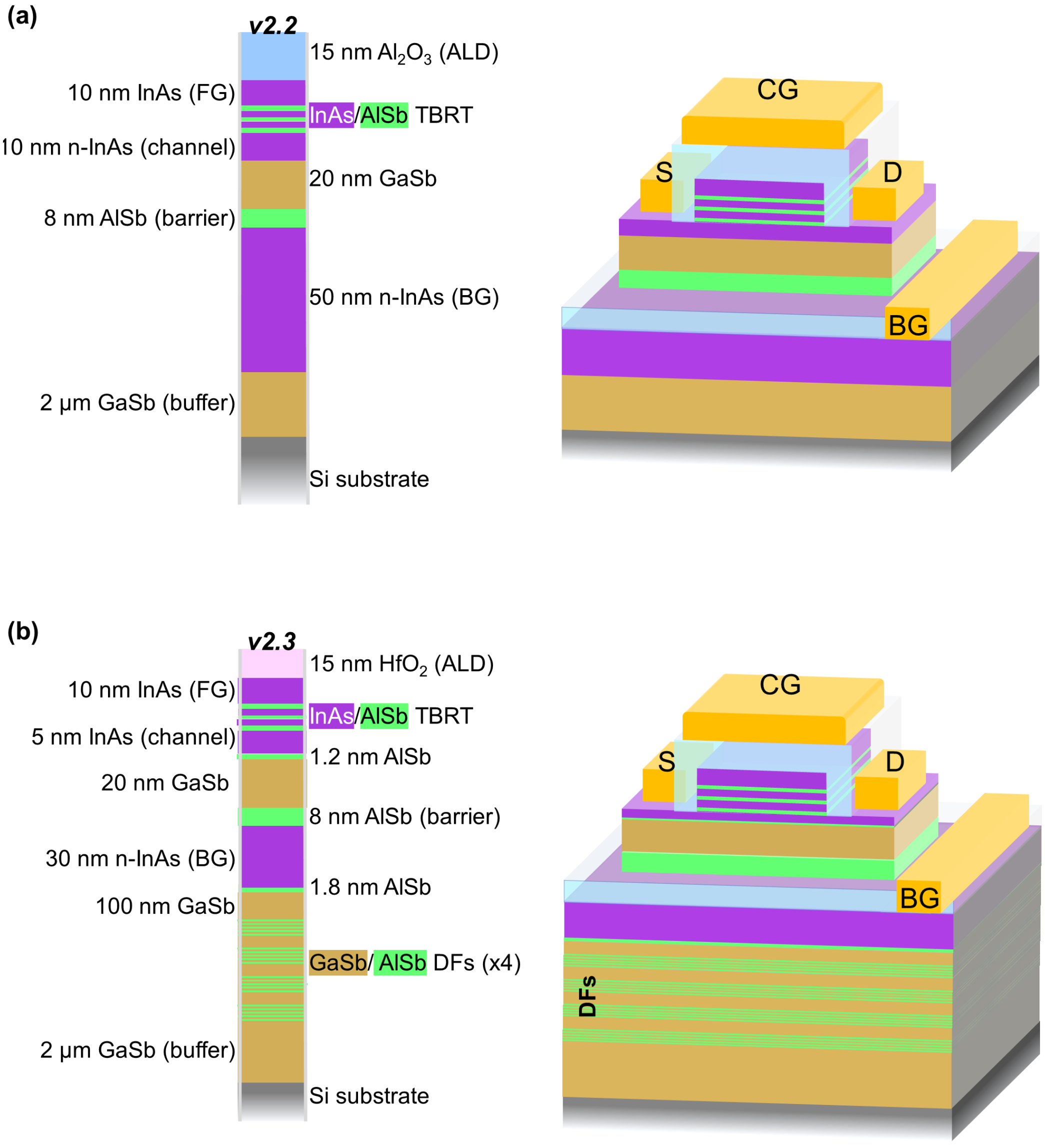


Figure 5.4: Schematic of the **(a)** *v2.2* and **(b)** *v2.3* ULTRA**RAM**™ designs for implementation on Si substrates (thicknesses not to scale).

5.2 UV Lithography Mask Design

The memory is fabricated on the MBE grown wafers using UV lithography. An optical lithography mask was designed which includes different sections corresponding to individual lithography steps. The mask for this process is spread out over two 4" quartz plates with chromium printed patterns. The mask layout was designed using AutoCAD [246] drawing software and the output file conversion was generated using KLayout [247].

The mask *Novel Nibble*[3] was designed to fabricate ULTRA**RAM**™ devices and arrays. In this section, an overview of the mask layout is presented in relation to the fabrication of 2×2 arrays, with the layouts for single devices and the backgate residing in Appendix C. It is important to note that these are presented separately to clarify the design in the reader's mind and are in fact on the same mask to be processed simultaneously.

Figure 5.5 depicts the full 3500 µm $\times$ 3500 µm unit cell repeated on the mask, in which each colour represents a different lithographic masking layer. The mask features single devices on the upper and lower areas of the cell. These device designs are replicated in 10 µm, 20 µm and 50 µm gate widths.

The central section of the unit cell contains 2×2 arrays which have four contacts each (two WL, two BL). These are arranged in reflected pairs in order to optimise space on the chip. The arrays are fabricated in 10 µm and 20 µm gate length where the array investigation places more emphasis on the architectural capabilities than performance metrics.

The unit cell utilises a U-shaped area for frontside backgate contact which wraps around the devices such that a contact can be made close to the target device (reducing contact resistance). This is an essential feature as the backgate is buried within the structure and is close to the surface of the crystal. The design also features TLM (transmission line measurement) bars to investigate contact resistances and optimisation line sets for calibration of the lithography process. The mask is designed to be used exclusively with positive photoresists.

5.2.1 Arrays

The 2×2 array design is similar to the single device design, albeit with four cells connected together and back-to-back D to BG connections to produce the architecture descried in the previous chapter. Here, we will present the detailed UVL design for the 2×2 array only, whereby the single device design is a singular version in which the S and D terminals are symmetric. Nevertheless, a full single device lithography description is provided in Appendix C.1, with a similar description for the BG lithography process in Appendix C.2.

[3] 'Nibble' or 'nybble' is a computing term for 4-bits, which is the size of the arrays on the mask.

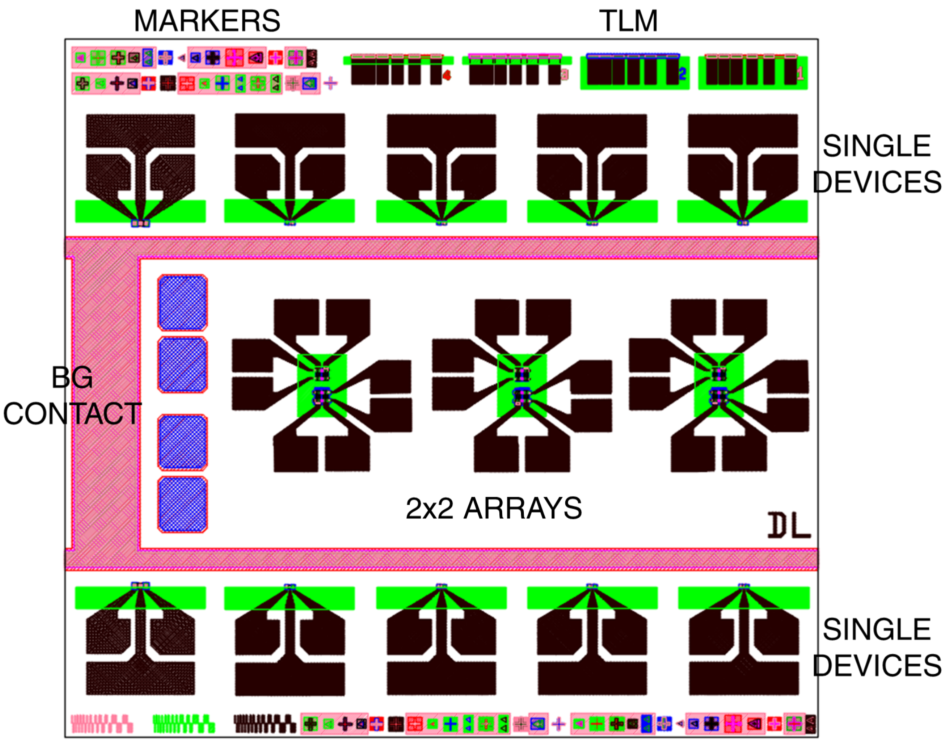


Figure 5.5: The unit cell overview of the *Novel Nibble* photomask. Single devices and 2×2 arrays of different feature sizes along with a U-shaped BG area, TLM bars (top right), alignment markers (top left and bottom right) and optimisation line sets (bottom left) are presented in the figure.

Mesa definition

For 2×2 arrays, the chrome features of the mesa mask are shaped to preserve enough material for the four connected devices, as shown in Figure 5.6**(a)**.

Source-drain fabrication

The source-drain fabrication masking layer [Fig.5.6**(b)**] forms a large continuous window across the mesa with four chrome rectangles inside. The material preserved within these chrome rectangles from the subsequent etch define the control gates of the nibble-sized array. The control gate size from the features yield 10 µm and 20 µm gate lengths. The outer edge of the open window is outside the mesa etch for alignment purposes.

Back gate fabrication

The array architecture features a shared drain and backgate contact for each pair of devices on the same BL. As such, the back gate masking layer has open windows on each BL between that which will later accommodate the drain contacts of the devices [Fig. 5.6**(c)**]. These resist windows enable an etch process to access the BG directly from the array which significantly improves the bit-density of the design.

Source-drain contact fabrication

The source drain contact masking layer [Fig. 5.6**(d)**] features open windows to define the contact terminals within the source-drain region of the devices, applied across the four devices which constitute the array. The key feature of this mask design is that the drain terminal openings are shared in pairs along the bitline and bridge across the back gate window. As a result, the drain and BG terminals of a pair of devices are set with one lithography window. This simplifies the process and increases scalability (and therefore bit-density) of the memory array as the number of interconnects per-bit is a key limiting factor in cell area reduction [249].

Control gate/wordline contacts

The CG/wordline contact mask design features open windows which overlap the S-D terminals [Fig.5.6**(e)**]. The control gates of pairs of devices within the same column are shared within this window. These are used to define the wordline (WL) of the memory array by application of metal coatings through the subsequent resist windows.

Oxide etching windows

Oxide etching windows formed from open glass are positioned on the mask in order to access the metal of the S and WL terminals deposited prior. As depicted in Figure 5.6**(f)**, the etching windows for the WLs are positioned away from the control gate regions of the devices to make way for subsequent BL deposition.

Lifting layer

Chrome features allow preserved resist to remain on the edge of the mesa in order to aid the path of conductive material from bond pad to the memory (detailed in 5.4.9). For arrays, this wraps around the entire mesa due to the increased number of contacts [Fig. 5.6**(g)**].

Final contact layer

The final contact layer features large windows for depositing metal bond pads with paths to the WLs and source terminals of the devices within the arrays. The window which is used for the connections of pairs of source terminals constitutes the array BLs, as they pass over the WLs of the array protected by the oxide layer below [Fig. 5.6(**h**)].

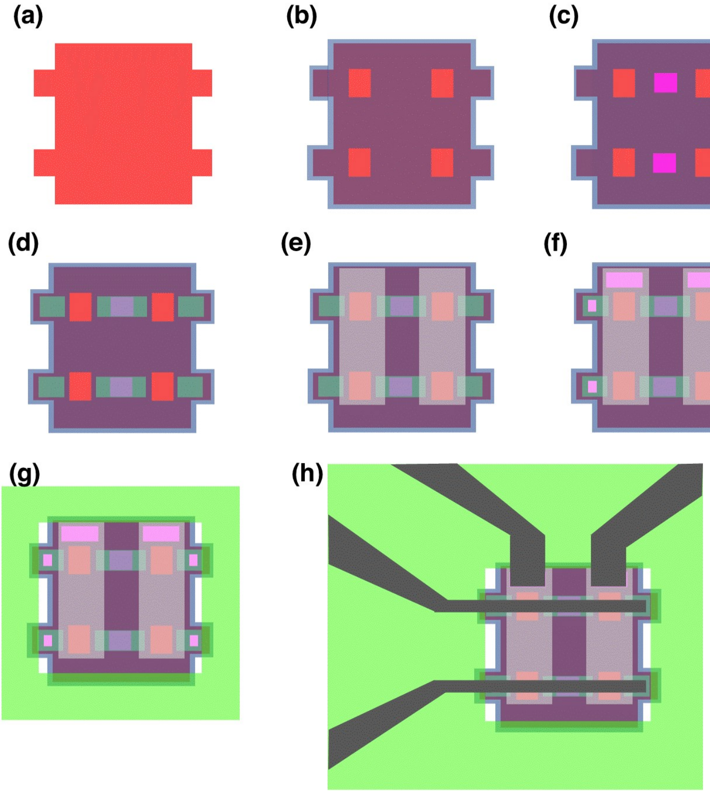


Figure 5.6: Detailed schematic of different mask sectors of *Novel Nibble*, each representing a UVL process step relating to 2×2 (4-bit) memory array fabrication. **(a)** Mesa definition. **(b)** Source-drain fabrication. **(c)** Backgate fabrication. **(d)** Source-drain contact fabrication. **(e)** Control gate/wordline contacts. **(f)** Oxide etching windows. **(g)** Lifting layer. **(h)** Final contact layer.

5.2.2 UV lithography process

The UV lithography process for producing the patterns from the masking layers involves using positive photoresists which are first spin coated onto the chip. Then, the resist is soft baked and exposed to UV light through the masking pattern before the exposed areas are developed using a suitable chemical solution.

The quality of the resist layer has a significant effect on process results and repeatability.

A detailed account of the most important factors and how they are considered in this work is provided in Appendix D.1 for the interested reader; which includes the processing details for resist application. The key outcome here are procedures for positive resist applications which allow for high-quality patterning for etching and metallisation processes.

5.3 Process Outline

In this section, the outline of the process flow will be described for fabrication of ULTRA**RAM**™ in both single device and array form for same-chip processing. Some individual samples presented in this thesis may have been processed in slight variations to this outline, the exact details of which can be found in the relevant appendices and are summarised at the end of this chapter. However, this outline of the process is true to the fundamentals of device fabrication for all samples.

5.3.1 Single devices

Figure 5.7**(a)** schematically represents the process flow for the fabrication of single memory devices. Each step of this process will be described in detail in due course, however this flow diagram serves as an outline. Beginning with the MBE-grown material, which includes the channel, TBRT layers and FG layer, we perform an etch to define the device mesa to electrically isolate devices/arrays from one another (Fig. 5.7**(a)**, step **I**). The mesa etch is executed using ICP dry-etching for high quality mesa floors with near-vertical sidewalls.

Another etch (using masking layer alignment with S-D masking pattern) is performed to define the source-drain regions of the memories (Fig. 5.7**(a)**, step **II**). The BG region is also etched at this stage such that subsequent etches for memory array BG access start out at the same etch depth. This etch must terminate in the channel layer and retain its full thickness for memory operation of NORMALLY-OFF (*v2.3*) designs. As such, a highly selective chemical etching process is employed and is detailed in 5.4.2.

ICP dry-etching is used with *in situ* reflectance monitoring to provide accurate etching to the n-InAs buried back gate layer away from the single device mesa. Next, the source-drain contacts are defined on the surface of the channel and also the BG using the LOR-3A/S1813 bilayer resist procedure from Appendix D.1.2. Thin titanium-gold contacts are applied by thermal evaporation or sputtering through the resist windows before lift-off (Fig. 5.7**(a)**, step **III**). Note that processing up to this point must be carried out expeditiously in order to minimise oxidation of the AlSb tunnelling barriers which are revealed following mesa and source-drain etches. When the sample is not undergoing a process it is kept under vacuum or in a nitrogen cabinet, and the process up to this point is completed on the same day.

A thin (10-15 nm) layer of Al_2O_3 (or HfO_2) is then deposited via thermal ALD (Fig. 5.7**(a)**, step **IV**). This layer serves three important purposes:

- Passivation of the devices to prevent further oxidation of the tunnelling barriers.
- Electrical isolation of the S-D contacts to produce overlapping CG contacts.
- High-k gate dielectric for the memory in which the top-InAs layer is the FG.

Metal contacts (Ti-Au) are then deposited onto the gate dielectric layer such that they encapsulate the gate stack of the device and overlap with the S-D contacts. This is required in order to provide an electric field across the channel for NORMALLY-OFF design operations (Fig. 5.7**(a)**, step **V**). An adhesion layer is then added to improve the bonding properties of the gold surface with subsequent layers, as discussed in 5.4.6. Then, using the Oxford Instruments PECVD 100 system, a SiO_2 layer is deposited to cover the entire sample. This step is non-essential for single device operation and is used to isolate BL from WL on the memory arrays. Although, the increased thickness in the bond pad areas will decrease the parasitic capacitances on the chip. A wet chemical etch is then used to remove both SiO_2 and Al_2O_3 selectively in order to reveal the source, drain, CG and BG contact regions (Fig. 5.7**(a)**, step **VI**).

A layer of hard-baked resist is applied; aligned with the edge of the device mesa on the contact side. As detailed later in 5.4.9, this step provides a path for the final contacts to stretch from the bond pads to the device contacts. Lastly, final contacts are applied which feature large bond pads for probing and wire-bonding with metal lines stretching from the pad, across the resist lifting layer and onto the devices to form the finished device. An SEM image of a 10 µm device is presented in the figure with red false colouring indicating the lifting layer.

5.3.2 Arrays

Figure 5.7**(b)** depicts a schematic process flow for memory arrays fabrication as a 2 × 1 formation for simplicity. The finished 2 × 2 array formation at 20 µm gate length is pictured in the SEM image below. The false colouring on the array indicates the extent of the WLs (yellow) and the etched access for the BL contacts to the S terminals (pink). The buried back-to-back D contacts,which are connected to the BG and isolated from the BL, are in the centre of each pair of devices (blue). Many steps are the same as for single devices which are processed on the same sample where single devices and arrays are processed simultaneously on the chip. In order to prevent repetition, we will focus on the key differences between the array and single device processes.

For memories with buried BG layers the mesa etch must be very precise. The source-drain etch leaves four FGs (2 × 2) remaining in which the entire array-region is etched down to the channel (Fig. 5.6**(b)** masking layer). This is carried out using the same wet-etch process as for single devices and is followed by a BG etch is repeated in the centre of the array between the device drain areas (Fig. 5.7**(b)**, step **II**, Fig. 5.6**(c)** masking layer). This process allows two drain contacts and a BG contact to be deposited in a single step (Fig. 5.7**(b)**, step **III**, Fig. 5.6**(d)** masking layer), which improves architecture bit-density.

The proceeding steps are similar to those of the single device design but applied across the four interconnected devices (Fig. 5.7**(b)**, step **III-VI**). In the final bond pad/contact deposition, the bitlines are arranged to intersect perpendicularly with the WL contacts (separated by the SiO_2 layer) to form the high-density crossbar array architecture.

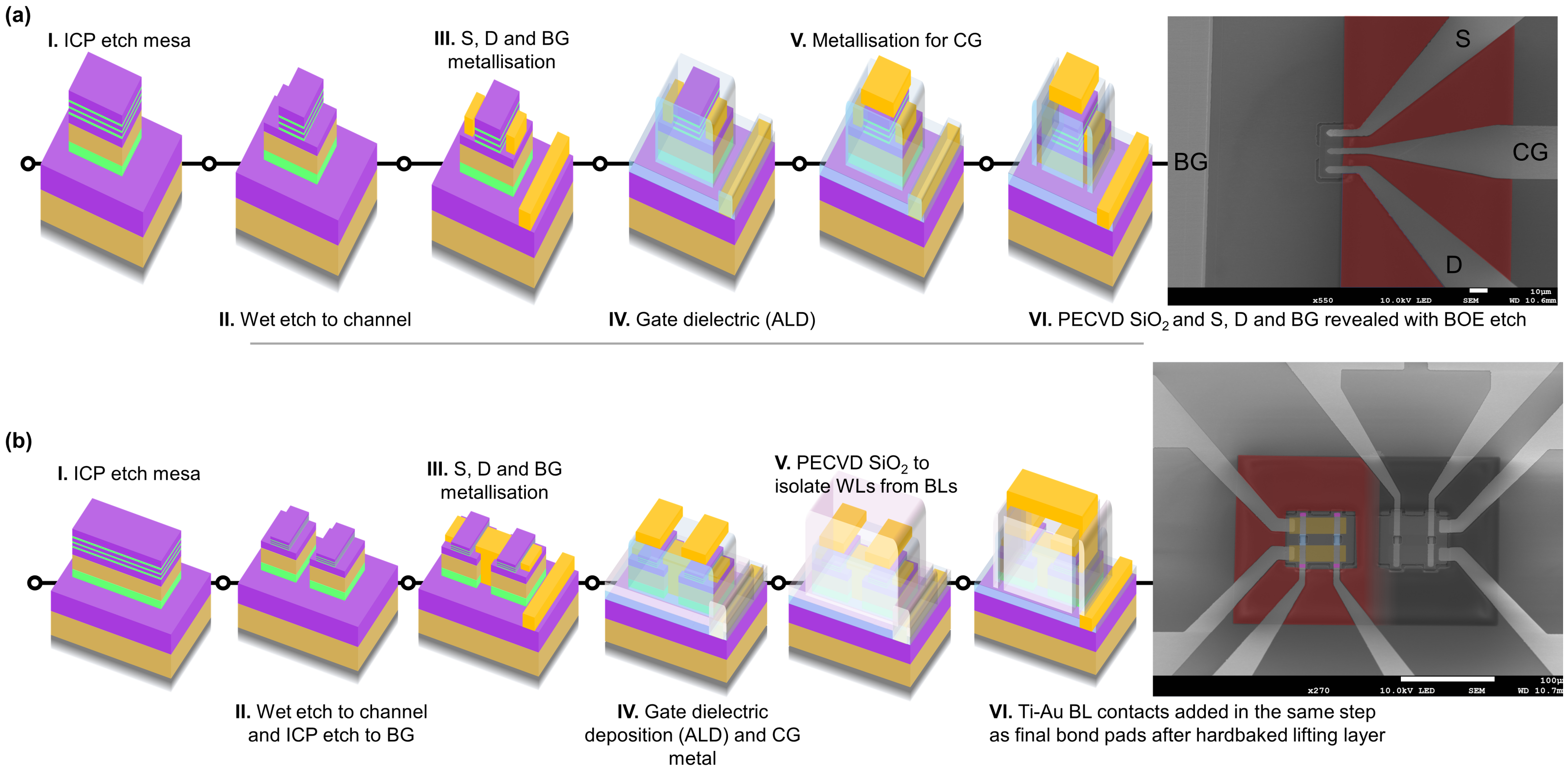

(a)
I. ICP etch mesa
II. Wet etch to channel
III. S, D and BG metallisation
IV. Gate dielectric (ALD)
V. Metallisation for CG
VI. PECVD SiO_2 and S, D and BG revealed with BOE etch
S
CG
BG
D
x550 10.0kV LED SEM 10μm WD 10.6mm
(b)
I. ICP etch mesa
II. Wet etch to channel and ICP etch to BG
III. S, D and BG metallisation
IV. Gate dielectric deposition (ALD) and CG metal
V. PECVD SiO_2 to isolate WLs from BLs
VI. Ti-Au BL contacts added in the same step as final bond pads after hardbaked lifting layer
x270 10.0kV LED SEM 100μ WD 10.7m

Figure 5.7: (a) Outline of process flow for single memory devices. Individual processes **I-VI** are described in detail within the text. **(b)** Schematic outline of process flow for a 2×1 memory arrays (half of a 2×2 array) to demonstrate array processing. Individual processes **I-VI** are described in detail within the text. SEM images of finished single devices and 2×2 memory arrays are shown at the end of the process flow.

5.4 Processing details

5.4.1 Mesa etch

The purpose of the mesa etch is to define the area of the memory device, isolating the devices from each other by etching the surrounding material to a certain layer. ICP dry etching with $Cl_2/BCl_3/Ar$ based plasmas produce a high-quality, smooth mesa floor with near-vertical sidewall etches. This gas mixture is found to be a very good choice for etching III-V semiconductors [250]. Plasma generated Cl* radicals penetrate into the InAs/AlSb/GaSb layers and chlorinate the Ga, In, Sb and As elements, resulting in the production of highly-volatile components to be removed from the surface. BCl_3 limits the chemical reaction of the chlorine plasma by lowering the amount of available reactive Cl provided by the Cl_2 gas, which improves process control [251]. Lastly, the introduction of argon into the reaction chamber greatly improves sidewall quality as argon plasma removal of etchant products by sputtering. Ar sputtering within the plasma chemistry also results in smoother surfaces due to a polishing effect when used at the appropriate forward power.

The gas mixture and pressure for the plasma was calibrated to produce the smoothest, cleanest etches for 6.1-Å semiconductor material layers at 25 W forward ICP power. The forward ICP power causes the plasma constituents to accelerate towards the sample and therefore controls the degree of ion bombardment on the surface. Thus, this parameter is a driving force in the etch rate and DC bias in the chamber.

In this work, the forward power and chamber pressure were adjusted from the 25 W calibration in order to change the etch rate. This choice is informed the level of process control required in the process. At low power, the polishing effect of the Ar is reduced such that Boron-containing etching products are not completely removed from the surface. This problem is solved by removing BCl_3 from the plasma. Indeed, contrary to the general rule, ICP etching of 6.1-Å materials produces a smoother surface without the inclusion of BCl_3, provided the forward power is less than 25 W[4] [251].

5.4.2 Source-drain etch

In order to etch through the memory structure to reveal the channel layer, a selective wet etch process was developed. A citric acid based solution is used to chemically etch InAs layers, however AlSb does not react with this solution such that the etch is halted upon arrival at the AlSb layer. Likewise, Microposit MF-319 developer, which is a TMAH[5] based solution, etches AlSb (and GaSb) but does not attack InAs. This allows selective etching of AlSb layers where the etch stops in the underlying InAs layer. Consequently, the memory structure can be etched layer-by-layer by

[4] 22 W used most often in this work.

[5] Tetramethylammonium hydroxide.

carrying out alternating dips in the two solutions to approach the channel through the InAs/AlSb TBRT structure with incredible precision. An alternative etchant which performs the same task as the TMAH is buffered oxide etchant (BOE) in 10:1 concentration. This is favoured for processing of later samples as the aggressive etching of the underlying GaSb through etch pits is reduced with this choice and is further improved with dilution (1:10 in de-ionised water). There are crucial details to consider when adopting the method on such thin material layers, which are provided in Appendix D.2.

5.4.3 Backgate etch

Access to the buried BG layer for both front-side contact and array contact was achieved using ICP dry etching. Details of gas mixtures and etching chemistries have been outlined previously in this section. For ULTRA**RAM**™ *v2.1*, 25 W forward power was used to access the substrate-level BG. For later memory iterations, forward power, pressure and gas mixtures were calibrated to provide the much needed etch control to stop the process within the thinner n-InAs buried BG layer, where the most successful and reproducible parameters were 22 W of forward power and 10 mTorr pressure in a Cl_2/Ar plasma.

The ability to terminate an etch in a certain layer within the crystal with nm-scale precision is a consequence of *in-situ* reflectance monitoring of the sample surface during the plasma etch. Additionally, the user must be aware of the reflectance dependence on etch depth in order to map the layers to the measured data on screen during the etch. For samples containing a wide range of materials and layer thicknesses, predicting the change of reflectance through the structure is not straightforward. A dedicated program, aptly named SimEtch[6], simulates the reflectance of the crystal structure during the etch at 670 nm (the wavelength of the Horiba® monitoring system installed on the ICP etch tool). This allows us to produce a simulated reflectance curve prior to performing an etch process. Then, the laser is aligned onto the area of the sample exposed to the plasma and the reflectance is measured during the etch with live-data available to the user. When the reflectance measured coincides with the simulated curve in relation to the desired etch depth, the process is abruptly terminated (manual jump) causing instant plasma shut-down which ceases the etch.

The results for the etching simulation for an ULTRA**RAM**™ *v2.3* wafer are shown in Figure 5.8**(a)**. Here, the reflectance is displayed as a function of etch depth from the wafer surface. The data obtained during the *in-situ* monitoring of the wafer are shown in Fig. 5.8**(b)** (black line). As the etch rate is vastly different depending on the material being removed, the InAs/AlSb layers take up a significantly larger portion of the plot. Nevertheless, the resemblance to the simulation is clear. The etch produces very smooth, high-quality features. The BG etch begins from the channel layer such that the reflectometry measurements also provide verification that the previous S-D chemical

[6] Developed by my colleague Dr. T. Wilson.

etch produced the desired result (Fig. 5.8**(b)**, red line).

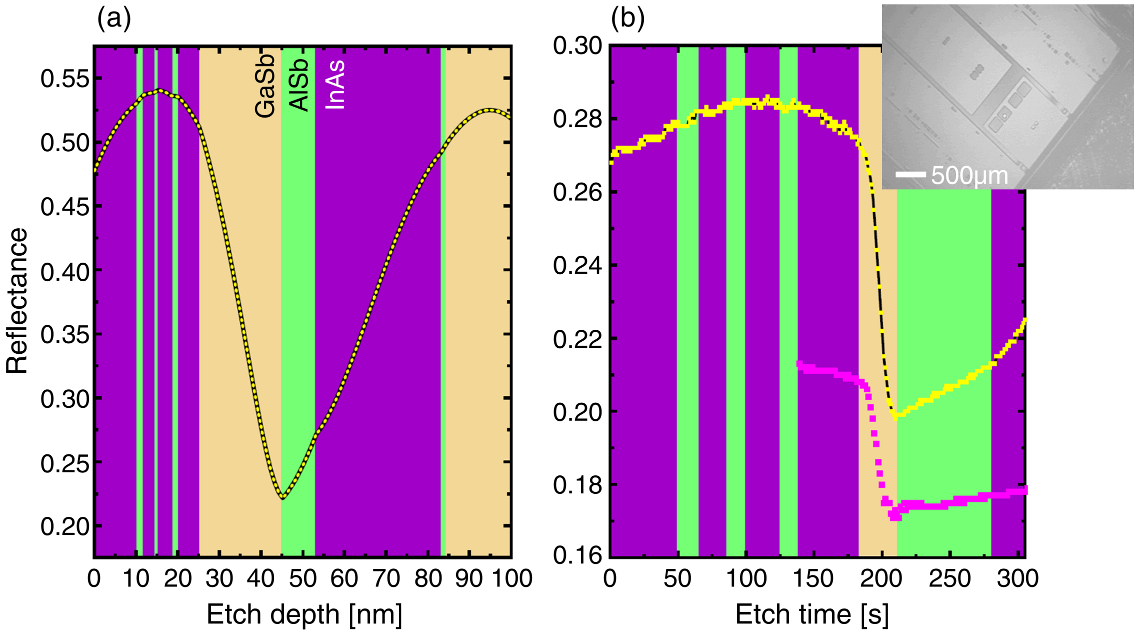


Figure 5.8: (a) SimEtch reflectance simulation data of the memory structure for *v2.3* designs. **(b)** Measured *in situ* reflectometry data during the ICP etch for an etch starting from the wafer surface (black line) and from the channel layer (red line). The time axis is offset to align the appropriate layers. The inset shows the microscope image used to align the laser for the reflectometry.

5.4.4 Source-drain contacts

Contacts for source and drain terminals are fabricated by sputtering or thermal evaporation of titanium and gold onto the surface of the sample through windows in a LOR-3A/S1813 bilayer resist (Appendix D.1.2) and onto the channel layer surface. When a metal and semiconductor are brought into contact, a potential barrier can be established at the interface at thermal equilibrium which prevents the flow of carriers across the junction. This phenomenon was first suggested by Walter Hans Schottky in 1938 as an explanation for the rectifying behaviour of such junctions and is therefore known as the Schottky barrier. The barrier height is a result of the difference between metal work function and the semiconductor electron affinity (for n-type) or ionisation energy (for p-type) [254]. Titanium and gold can be employed for both InAs and GaSb contacts due to the strong Fermi-pinning nature of metal/III-V interfaces. This provides a low Schottky barrier height (ϕ_{Bn}) for electrons at the metal InAs (or InGaAs) heterojunction and for holes (ϕ_{Bp}) at the metal/GaSb heterojunction as depicted in Fig. 5.9 [255, 256], which are suitable for n-MOSFET and p-MOSFET operation respectively (*i.e.* low-resistance, ohmic channel contacts). In fact, InAs has a peculiar tendency to adjust its energy bands in such a way that the Fermi level (E_F) becomes pinned above the conduction band minimum (CBM). This occurs due to formation of an electron accumulation layer in the near-interface region where surface-states pin the Fermi energy [207].

Gold has an extremely small resistivity and is chemically stable in atmospheric conditions. The gold surface also provides a high-quality bonding surface for device packaging and is therefore an excellent choice of metal for device processing in a research setting[7]. Gold has poor adhesion properties when interfaced with many non-metallic materials, including III-V semiconductors. The best possible way of ensuring adhesion of deposited material layers is to ensure a reaction with the functional groups on the material surface; in particular, those which provide stable metal-oxide interfaces [257]. Since gold is a noble metal, this is rarely the case and there is an absence of chemical bonds between the gold and the substrate which results in poor adhesion properties.

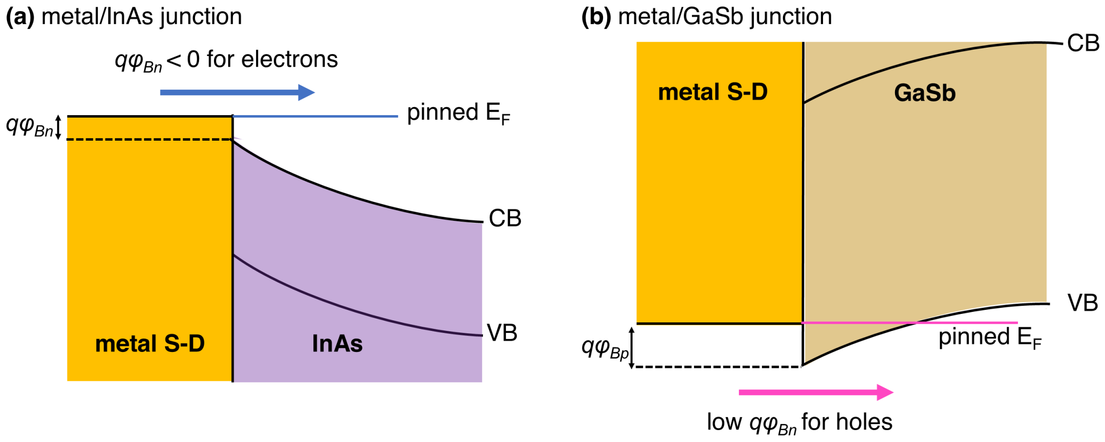


Figure 5.9: Energy band diagram for **(a)** metal/InAs junction with the Fermi energy (E_F) pinned above the InAs CB and **(b)** metal/GaSb junction with a small Schottky barrier (ϕ_{Bp}) due to Fermi level pinning.

There are a few materials which can help resolve the gold adhesion problem. Most notable are the common oxide forming metals such as titanium or aluminium which can create oxide interfaces more easily. As such, Ti and Al adhere exceptionally well to dielectric and semiconductor surfaces [257]. A thin (few nm) layer of Ti is deposited prior to gold, which acts as an adhesion promoter. Ti has a work-function of 4.33 eV, which is acceptable for InAs (and InGaAs) and GaSb interfaces [258]. The gold deposited onto the thin metallic titanium behaves more like a very thin Ti/Au alloy, which is known to have an exceptional bonding strength. It is extremely important to maintain vacuum between Ti and Au depositions to prevent oxidation of the thin Ti layer which would disrupt Ti-Au alloy formation.

[7] Copper (Cu) and aluminium (Al) are often used for economic reasons and Au is not compatible with commercial CMOS processing due to contamination issues.

5.4.5 Gate dielectric deposition

The gate dielectric is deposited over the entire sample. This is carried out by ALD, where Al_2O_3 is the initial material of choice, and HfO_2 was introduced later in an effort to increase FG capacitance. Alternative deposition techniques and materials were tested to use as the gate dielectric (PECVD of SiO_2 and e-beam evaporation of Al_2O_3) however these do not provide a high-quality leakage-free gate dielectric at the required film thicknesses[8]. A layer of Al_2O_3 is deposited on the sample by cycling ($\sim$ 1 Å/cycle) of TMA/H_2O in the thermal ALD process. The chosen ALD recipe for gate dielectric deposition is presented later in Table 5.2 where alterations to material choice and growth temperature were motivated by potential improvements to the memory window (capacitance increase) and interface quality respectively. The details of this common thermal ALD method have been discussed previously in subsection 3.3.8. HfO_2 gate dielectrics were deposited also using thermal ALD under similar conditions, where the Al-precursor used for Al_2O_3 is substituted for TDMA-Hf[9].

As the chemical cycle involves bonding to hydroxyl groups, ALD deposited films have very good adhesion properties, even on noble metals [259]. The resulting film is exceptionally uniform, coating the entire surface (including sidewall features), producing a dielectric layer that is highly resistive, pinhole-free and high-k.

As well as the gate dielectric which isolates the device FG, the ALD layer provides isolation between the S-D and CG contacts for the S-D/CG overlap design. It is important that the separation distance between the overlapping contacts is small such that there is sufficient electric field to activate the NORMALLY-OFF channel for readout operation. The separating layer must also be highly resistive to prevent current leakage between the device terminals. Therefore, ALD of metal oxides is a perfect choice for this purpose.

ALD-deposited Al_2O_3 provides outstanding passivation of III-V semiconductors. The early stages of the ALD-process on III-Vs cause self-cleaning and native oxide removal on the surface. This phenomena is based on the interaction of $Al(CH_3)_3$ (TMA precursor) with native oxides of (In)GaAs surfaces. This leads to most of the surface arsenic oxides and a significant portion the indium/gallium oxide being consumed. It seems that just a half cycle of Al_2O_3 could passivate the majority of the defects [260], but it is known that dissociative chemisorption of $Al(CH_3)_3$ leads to the formation of an ordered monolayer of dimethyl aluminium. This gives rise to the formation of metal–metal bonds that pin the Fermi level, so dosing with H_2O (*i.e.* a completed cycle) is required to passivate the surface and unpin the Fermi level by inserting O atoms in the metallic bond [261]. Indeed, studies of MOS-capacitors with similar InAs/Al_2O_3 interfaces exhibit electrical properties which confirm that the Fermi level is not pinned [262]. This is crucial, as Fermi level pinning at this interface would otherwise disrupt the NORMALLY-OFF channel operation, as the Fermi level would be pinned above the In(Ga)As QW_{CH} ground state at zero bias. Note that a similar passivation

[8] Large gate leakage currents are observed directly in failed memory samples using these techniques.

[9] Tetrakis(dimethylamido)hafnium(IV).

process has been observed for InAs interfaces with ALD-HfO_2 chemistries, where again the metal-methyl precursor consumes surface oxides to produce a metal-oxide layer prior to introduction of the second precursor (H_2O) [263].

5.4.6 Control gate/wordline contacts

The CG/WL contacts are fabricated from sputtered or thermally evaporated Ti-Au in a similar fashion to the S-D contacts as previously described, albeit with one important difference: Before ending the process, an addition ultrathin (around 3 nm) Ti layer is added to produce Ti-Au-Ti contacts. When the vacuum is broken, the thin uppermost Ti layer is exposed to air and reacts to become titanium oxide. Later, this acts as an adhesion promoter between the CG/WL contacts and the SiO_2 isolation layer(s). The adhesion chemistry here is the same as for the initial Ti-Au contacts, as outlined previously. Without the final Ti layer, catastrophic process faults occur during the oxide etching step due to adhesion failures, which are discussed in detail in Appendix D.4. An alternative, equally effective technique used on later iterations is to add 2 nm (20 cycles) of ALD-Al_2O_3 after Ti-Au deposition to act as the bonding promoter for subsequent layers.

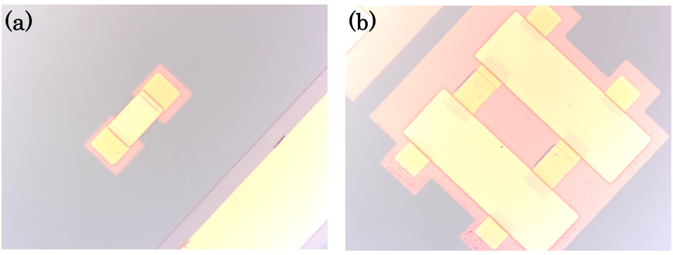


Figure 5.10: Optical microscope images of ULTRA**RAM**™ memories after the overlapping CG/WL metallisation for **(a)** single devices and **(b)** 2 × 2 arrays.

5.4.7 Oxide isolation

Silicon dioxide (SiO_2) is deposited everywhere on the sample using the Oxford Instruments PECVD tool. The oxide provides isolation between WL and BL contacts within the memory array architecture. Deposition took place at 100 °C at a rate of 60 nm/min. Film thickness of the depositions was around 120 nm (100 s deposition time[10]) for all samples in this work. The deposition rate was investigated by depositing the oxide on multiple clean Si samples for various deposition times and measuring thickness using an optical ellipsometry technique.

[10] The starting thickness at 0 seconds is non-zero due to material deposition in the multiple plasma-striking steps prior to the main deposition.

The deposition thickness on the memory sample can be approximated during the process with the naked eye. Optical interference results in a distinct variation in sample colour depending on material thickness. Thus, the deposition thickness can be quickly inferred from colour checking the chip post-process [264]. In this work, all samples emerged from the PECVD reaction chamber with dark blue surfaces, corresponding to an oxide thickness of 120-130 nm.

5.4.8 Oxide etching

To regain access to the CG, S and D contacts on the single memory devices and the WL and S contacts on the memory arrays, the oxide layer(s) in these regions are chemically etched through exposed UVL windows. In the CG/WL regions of samples A and B, there is a SiO_2 layer and thin TiO_2 (for adhesion purposes) layer separating the gold contact from the sample surface. Whereas, in the S-D regions, there is a SiO_2 layer and ALD gate dielectric above the gold contacts.

BOE 10:1[11], a diluted HF acid solution containing a buffering agent (NH_4F), etches all the materials required here at the following rates:

- TiO_2 = 90 nm/min [266] for dilute HF (10%).
- ALD-Al_2O_3 = 30 nm/min [267] for 20:1 BOE.
- PECVD-SiO_2 = 490 nm/min [268] for 5:1 BOE.

Etching of these material layers was tested and calibrated on multiple Si samples with the same material depositions as the memory chips (15 nm Al_2O_3 and 100 nm SiO_2), the details of which are provided in Appendix D.4. The results demonstrate that a 60 s dip time is sufficient for the material layers used in most samples (A-C). When introducing the hafnia gate dielectric to the technology, the 60 s dip is insufficient to remove the layer. As the HfO_2 etch rate in HF is slow (<1 nm/min etch rate), a 60 s HF dip is first used to remove the SiO_2, followed by an ICP dry etch by $BCl_3/Cl_2/Ar$ plasma at 25 W forward power and 4 mTorr chamber pressure, where *in situ* reflectometry is used to confirm the removal of the hafnia layer. The etch rate of the material under these conditions is 10 nm/min, where the underlying titanium-gold terminals prevent damage to the III-V layers of the device.

5.4.9 Contact lifting layer

A hard-baked photoresist lifting layer was patterned around the edges of arrays and on a single edge of the individual memory devices. The purpose of this layer is to cover the etched outer edges of the device and array mesas. Due to the outward incline of the photoresist, these regions

[11] A common commercial pre-prepared solution made by SigmaAldrich was used here [265].

allow continuous metal contacts to form across them. This enables metal bond-pad contacts to connect to the device terminals smoothly, where the metal deposition is not disrupted by any vertical features.

5.4.10 Final metal contact deposition

Finally, metal-contacts are formed on the sample which provide bonding pads for later packaging wires which connect to the CG and S-D terminals of single devices and WL and BL contacts of the arrays. The metal contacts traverse the contact lifting layer (photoresist) to produce smooth, continuous conducting paths from device to bond pad. The area of contact on the memories is in the location of previously etched oxide material.

The contacts are sputtered or thermally evaporated using titanium-gold in the same fashion to previous metal depositions. Here, the initial Ti layer acts as an adhesion promoter for the SiO_2 sample surface. A thick gold layer (>150 nm) is added using the thermal evaporation method[12] (subsection 3.3.5). The anisotropic deposition of material from evaporation enables application of thicker metal layers onto the sample without compromising the resist lift-off procedure. Increased metal thickness reduces contact resistance and greatly improves probability of successful wire-bonding. Further details on this procedure are available in Appendix D.3.

5.5 Sample Summary

The differences in process and memory design between samples in this work are summarised in Table 5.2. The overall design remains fairly similar as the Si growth integration, channel etching uniformity and ALD processes are developed and improved with each attempt. An n-type InAs NORMALLY-ON channel is used in samples A-C to simplify the structure whilst introducing the ALD gate dielectric and Si substrate.

The choice of selective etchants for revealing the channel layer changed through sample iterations. The etching of the AlSb TBRT barriers was found to give a smoother etch morphology with BOE compared to MF-319 (samples A-C) and was improved further by aqueous dilution (sample D). Titanium-gold contacts fabricated by sputtering were found to have a higher contact resistance than thermally evaporated ones, so later samples (B-D) use evaporated contacts exclusively. The final memory sample (D) features a HfO_2 gate dielectric in an attempt to widen the memory window by increasing floating gate capacitance, as outlined in the previous chapter.

[12] If using sputtering, the gold layer is thickened further by thermal evaporation of gold only.

Table 5.2: Summary of processing designs for different iterations of ULTRA**RAM**™ memory.

	Sample A	**Sample B**	**Sample C**	**Sample D**
Substrate	GaAs	Si	Si	Si
Wafer ID	XPH1422	XPH1452	XPH1452	XPH1642
Tech. iteration	*v2.1*	*v2.2*	*v2.2*	*v2.3*
Channel layer	12 nm n-InAs	10 nm n-InAs	10 nm n-InAs	5 nm InAs
Channel operation	NORMALLY-ON	NORMALLY-ON	NORMALLY-ON	NORMALLY-OFF
Channel etching	Citric/MF-319	Citric/BOE	Citric/BOE	Citric/dilute BOE
Gate dielectric	Al_2O_3	Al_2O_3*	Al_2O_3	HfO_2
ALD thickness	15 nm	15 nm*	15 nm	15 nm
ALD number of cycles	150	150	150	130
ALD temperature	150° C	150° C	200° C	150° C
Metallisation technique	Sputter	Evap. (Al gate)	Evap.	Evap.
CG adhesion layer	Ti	Ti	Al_2O_3	Al_2O_3

* indicates the gate dielectric was unintentionally etched in this sample.

Chapter 6

Electrical Measurements Results and Discussion

The room-temperature electrical characteristics of fabricated ULTRA**RAM**™ single devices and arrays are presented and discussed in this chapter. Each sample (A, B, C and D) has slightly different layer design and fabrication steps. These details are presented in Section 5.5.

6.1 Sample A

Sample A is a memory chip where the III-V epilayers are grown on a GaAs substrate, a growth design known as ULTRA**RAM**™ *v2.1*. The main differences between this design and the one used in initial prototypes (*v1.0* [216]) are the reversal of the tunnelling direction to incorporate the ALD gate dielectric and thinner channel layers to bring the BG layer closer to the FG for improved gate response.

6.1.1 Channel measurements

Transmission-line measurement (TLM) is a technique which allows the contact resistance and semiconductor sheet resistance to be determined through simple current-voltage (I-V) measurement when considering the geometry of the TLM structure, which is depicted in Fig. 6.1. If we consider the semiconductor sheet the TLM contacts are placed upon as a simple resistor, the total resistance, R_T, measured between two contacts is

$$R_T = 2R_m + 2R_c + R_s \ , \tag{6.1}$$

where R_m is metal contact resistance, R_c is metal-semiconductor interface resistance and R_s is the semiconductor sheet resistance. For our titanium-gold contacts used here, R_m can be neglected as $R_m \ll R_c$, such that contact resistance is typically dominated by the metal-semiconductor interface properties.

The resistance of the semiconductor sheet is proportional to the distance between the two contact bars, described by

$$R_s = \rho\left(\frac{d_{\text{i-j}}}{A}\right) \, ,$$

where ρ is the resistivity, $d_{\text{i-j}}$ is the distance between the i'th and j'th TLM bars and A is cross sectional area, calculated from the width of the TLM bar, Z, and the thickness of the semiconducting layer, t. This results in following for semiconductor resistance:

$$R_s = \frac{\rho}{t} \times \frac{d_{\text{i-j}}}{Z} = R_{sheet}\left(\frac{d_{\text{i-j}}}{Z}\right) \, , \tag{6.2}$$

where R_{sheet} is the sheet resistance ($\Omega/\square$). The resulting equation for the total measured resistance is

$$R_T = R_{sheet}\left(\frac{d_{\text{i-j}}}{Z}\right) + 2R_c \tag{6.3}$$

This allows R_c and R_{sheet} to be determined from TLM resistances by plotting resistance between pairs of contacts against their separation distance, $d_{\text{i-j}}$. The result should be linear, where the gradient and y-axis intercept are used to determine sheet and contact resistance respectively.

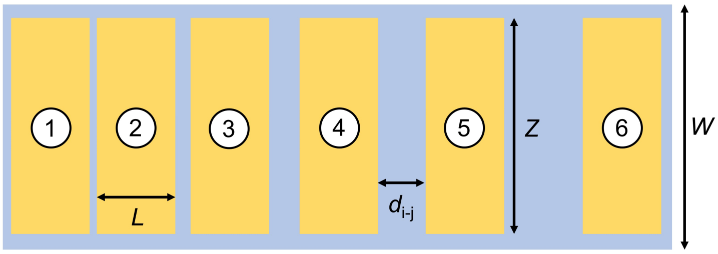


Figure 6.1: Structure of TLM bars on the mask for resistance measurements. Gold rectangles indicate the titanium-gold TLM contact bars on top of the semiconductor layer.

The analysis described above assumes a linear I-V characteristic (Ohmic resistances) for all values. Surprisingly, the TLM for the n-InAs channel layer for sample A exhibited a diode-like response with high resistances at low bias, an example of which is shown in Fig. 6.2. Moreover, there is large variation across the chip and the TLM measurements show no distance dependence, indicating that the dominant resistance is at the metal-semiconductor interface where there is a Schottky barrier as described in 5.4.4. This is unexpected for this material system, and persisted in the TLM I-V characteristic for the n-InAs BG layer also. Through careful analysis of contact resistances and EDX mapping of the contact areas, it was concluded that this issue was a result of poor titanium-gold sputtering quality, where the thin titanium layer becomes oxidised to produce an energy barrier at the interface. Later samples used thermal evaporation to deposit metals which resulted in low-resistance Ohmic contacts as expected. Indeed, the high resistance device measurements for this sample are a result of metal deposition quality alone.

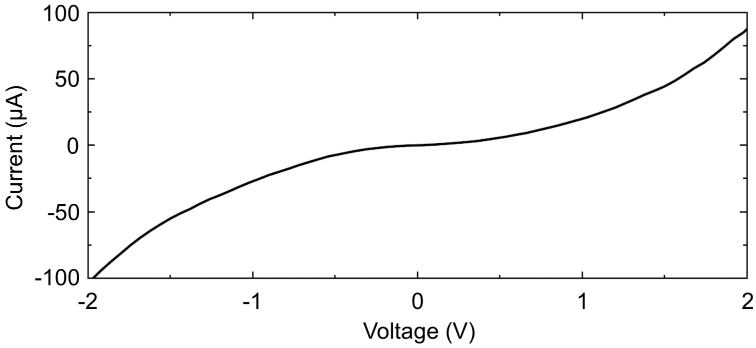


Figure 6.2: I-V characteristic for on-chip TLM of Sample A with a pad spacing of 135 µm.

6.1.2 Program and erase at low bias

The S-D current-voltage ($I_{\text{S-D}}$-$V_{\text{S-D}}$) relation[1] is shown for the programmed (0) and erased (1) states in Fig. 6.3 by red and black lines respectively. The device presented here is contained within a 2×2 array with a 20 µm gate length. In the array architecture, the D contact is accessed through the shared BG terminal in the centre of the array. The program cycle (P) consisted of a -2.5 V, 500 µs pulse across the S-D terminals with the CG terminal grounded, whilst the erase cycle (E) used a +2.5 V pulse under otherwise similar conditions. In the programmed state (Fig. 6.3, red line), the I-V shifts to higher resistance relative to the erased state (black line) as the negative potential in the FG depletes the n-InAs channel. A S-D voltage with CG grounded is required to perform P and E cycles in these devices. This is because a voltage across the channel is required to overcome the energy barrier at the metal-channel interface in order to apply potential across the gate stack.

The observation of memory switching at ±2.5 V with a clear state contrast confirms the ULTRA**RAM**™ low-energy operation as outlined previously. This corresponds to a 10^2 and 10^3 reduction in switching energy per unit area in comparison to DRAM and NAND flash respectively [75]. The inclusion of the ALD-Al_2O_3 gate dielectric eliminates CG leakage entirely ($> 10^{11}$ Ω), providing full FG isolation. The speed of the P/E cycle is noteworthy: the 500 µs pulse duration used here represents a 2000-fold performance improvement over the initial memory prototypes [216]. This source of the improvement is likely to be from the inclusion of the ALD-Al_2O_3 gate dielectric which eliminates the leakage current through the gate stack. As outlined in 2.2.1, quantum tunnelling is of the sub-ps scale and switching speed in floating-gate memories is limited by its *RC* time constant. Thus, ULTRA**RAM**™ is subject to Dennard's scaling law [7] and should receive significant

[1] At zero CG bias.

performance upgrades from shrinking the lithography to modern-day feature sizes. Indeed, an ULTRA**RAM**™ 20 nm feature size device corresponds to sub-ns switching speed following ideal scaling. This is significantly faster than DRAM and comparable to SRAM. Moreover, this observation is consistent with the simulated performance in Section 4.7, although rigorous testing on scaled devices is required to confirm this.

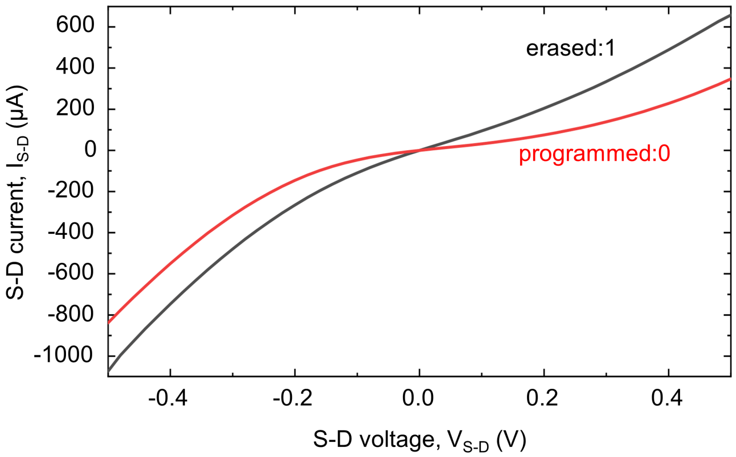


Figure 6.3: I-V sweep of S-D voltage with measured S-D current in the absence of CG bias after an erase cycle of 2.5 V, 500 μs duration (black line) and a program cycle of -2.5 V, 500 μs duration (red line). Figure reproduced from [233].

Half-voltage architecture

P/E cycles carried out on the device at ±2.5 V on across the BL (S) to BG (D) on the array only program the target device with its WL grounded. Other devices on the array with a shared BG (S) are undisturbed as their WL (CG) is floating such that no potential is across the gate stack. Previously, in Section 4.7, a half-voltage architecture for P/E cycles in the array is described in which half the voltage for the P/E cycle is applied to the WL (CG) and the other half is applied to the BL[2] (S), selecting column and row to P or E the target device without disturbing surrounding cells.

Architecture testing was carried out on a single 20 μm gate length device in a 4 bit array by direct probe station measurements, the results of which are shown in Fig. 6.4. As predicted, the P and E is successfully performed on the device with ±1.25 V on S and CG to form the 2.5 V potential across the gate stack. Fig. 6.4 demonstrates this for 5 manual P/E cycles. Then, disturbance tests consisting of an uninterrupted ±1.25 V bias were applied separately across each terminal for 120 s, which investigates the disturbance voltage the cell would encounter within the half-voltage architecture. This is an equivalent disturbance of $> 10^5$ P/E cycles, as the P/E speed of the devices is 500 μs. The half-voltages slightly shift the position of the memory logic state.

[2] Of opposite polarity.

Moreover, the direction of any significant shift tends to be in accordance with the loss or gain of electrons in the FG during the test. Importantly, the perturbation of the states is not sufficient to disturb the logic in a way that the memory logic is lost entirely (Fig. 6.4). This experimentally demonstrates that the disturb rate for the proposed architecture is very small, as predicted by simulations of the device physics from 4.8.

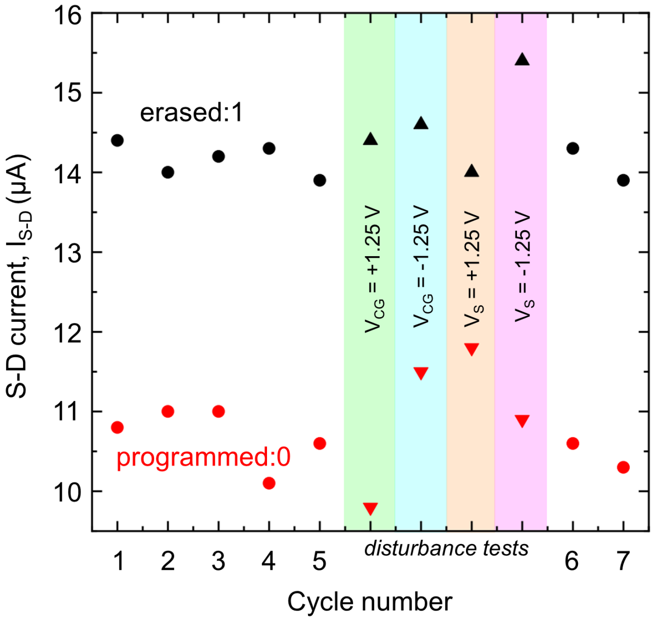


Figure 6.4: Half-voltage manual cycling of ULTRA**RAM**™ using ±1.25 V voltages simultaneously on CG and S terminals (circles) before testing disturbance from individual half-voltages (triangles) applied for 120 s. A 0.2 V S-D bias was used for readout.

6.1.3 Retention

Memory retention (*i.e.* non-volatility) is tested on a device of 20 μm gate length within a 2×2 array, where the P (0) and E (1) states are set using S-D voltages of -2.5 V and +2.5 V pulses respectively, with the WL (CG) grounded. The memory state is read over 8×10^4 seconds (>22 hours) by measuring the S-D current ($I_{\text{S-D}}$) at a 0.5 V S-D bias ($V_{\text{S-D}}$) and zero CG bias. The non-volatility is demonstrated in this test in Fig. 6.5, where the P state (red, 0) and E state (black, 1) are read 8×10^4 times during the retention test. The stability of the FG charge retention is a result of the 2.1 eV barriers from the InAs/AlSb heterostructures on the channel-side of the FG, and the 3.1 eV barrier from the InAs/Al_2O_3 interface on the CG side [205]. This test confirms the electrons tunnelling into the FG via the low-voltage TBRT mechanism can be robustly stored to form non-volatile logic states. Moreover, a non-destructive, low-disturb readout of the memory is demonstrated.

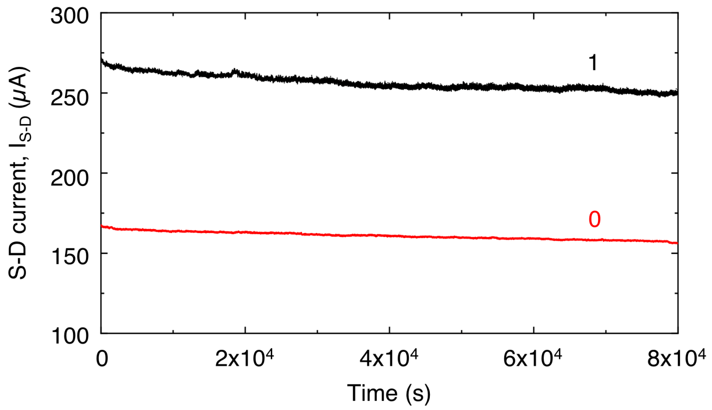


Figure 6.5: Retention data for a 20-μm gate length cell in a 2×2 memory array. Memory logic is programmed (0) and erased (1) with 500 μs pulses of -2.5 and +2.5 V on the S terminal, respectively. Read-out current is measured with a 0.5 V source voltage every second. Figure reproduced from [233].

6.1.4 Endurance

Floating gate memories (*i.e.* flash) suffer from poor endurance[3], such that wear-levelling and bad-block management is required to prolong their lifetime [67]. The mechanisms which cause device failure in flash technologies have been discussed previously in 2.2.1, which are mostly related to the high voltage required for the P/E cycles. A memory technology for RAM applications requires superior endurance properties to ensure a good lifespan, as individual memory cells are programmed and erased with each computational operation. Memory cells from sample A withstood 10^6 P/E switching cycles (P-read-E-read), maintaining a clear 0/1 state contrast with each cycle [Fig. 6.6**(a)**]. The switching cycles were carried out at a rate of 200 cycles per minute, with 5 ms P/E pulses of $\pm$2.5 V, except for the blue shaded region, where the pulse duration was reduced to 500 μs. This reduces the 0/1 contrast [Fig. 6.6**(b)**], which is due to the *RC* constant of the device feature size (*i.e.* the gate-stack potential does not reach 2.5 V within the pulse). The tunnelling mechanism is intrinsically extremely fast, as detailed in 4.7.1 [218].

In this first-time test, endurance is demonstrated to be at least an order of magnitude improvement compared to flash memory [75]. However, there is movement of the 0/1 window throughout the duration of the test. The reason for this is not known, however the most likely reason is that it is related to device fabrication. The issues with S-D contact deposition produce inconsistent contact with the device channel. Moreover, the channel itself shows significant etch pitting from the selective etchants used previously. Details of the etch and AFM measurements identifying the pits can be found in Appendix D.2. It is thought that the sub-optimal processes identified in this sample cause a highly inconsistent channel contact which is sensitive to temperature

[3] Degradation due to many P/E cycles.

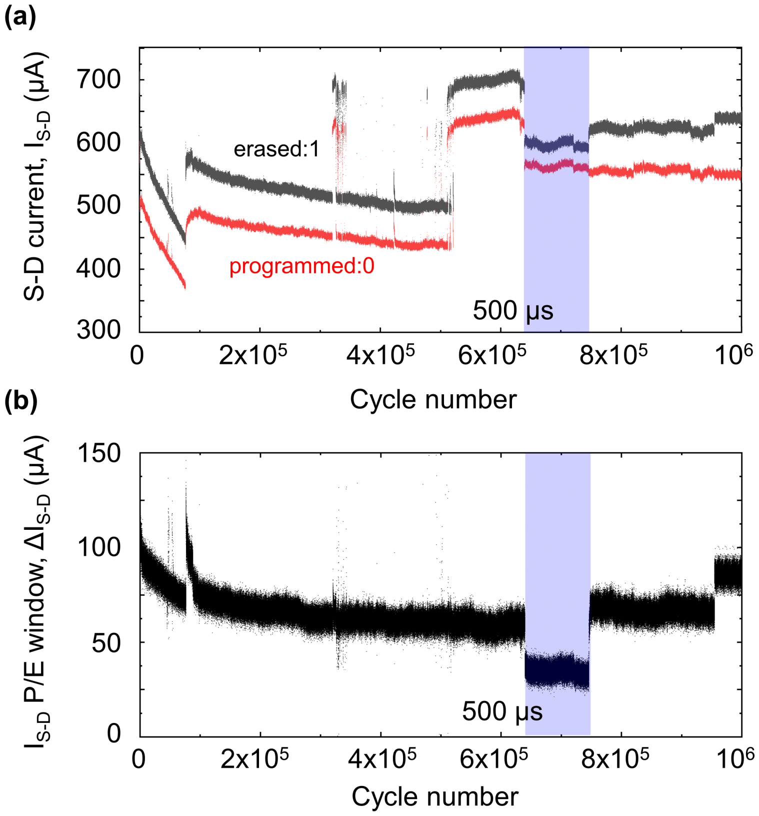


Figure 6.6: Endurance data for a memory cell in an array (sample A). Readout is performed with a 0.5 V S-D voltage every second. **(a)** S-D current after a +2.5 V erase cycle (grey), and a -2.5 V program cycle (red). Pulse duration was set to 5 ms, except for those data points with blue shading where a 500 µs pulse duration was used. **(b)** S-D current difference calculated by subtracting erase and program current from consecutive cycles. Figure reproduced from [233].

or external vibrations. This explanation is supported by clearer $I_{\text{S-D}}$ data on Sample C, discussed in Section 6.3. Fluctuations in the offset of the overall $I_{\text{S-D}}$ aside, the memory state is observed with similar current difference between 0 and 1 states: $\Delta I_{\text{S-D}}$ persists for 10^6 cycles, as shown in Fig. 6.6**(b)**, with P and E states following each other.

6.1.5 Uniformity

The principles of ULTRA**RAM**™ have been demonstrated on GaAs substrates with promising results for both performance and reliability and a compact architecture compatibility has been identified. However, process uniformity and device stability require improvement. The conducting n-InAs NORMALLY-ON channel provides a relatively small current difference between 0 and 1 states, as there is no threshold voltage (V_T) channel response to implement a readout mechanism similar to flash memory (2.2.1) making array operation difficult. However, the small memory window ($\Delta I_{\text{S-D}}$)

in sample A is not an indication of logic state weakness, but rather due to the simplicity of the channel construction at this stage of the memory development. In later samples, an effort is made to improve process uniformity and develop a channel design with a V_T to improve state contrast similar to those described in Section 4.6. First, the memory design will be implemented on a silicon substrate with improved processing before implementing more complex channel designs on the new substrate material.

6.2 Sample B

Sample B is a first attempt of a memory process on GaSb/Si substrate. The MBE-grown layers for memory operation are otherwise unchanged from those of sample A. In fabrication, the gate metal was substituted for Al in an attempt to improve the interface quality. However, the later HF-based etchants caused damage to the Al and compromised the Al_2O_3 gate dielectric residing beneath the gate metal. Although this destroys the memory retention function of the memory due to FG-leakage, the absence of the gate dielectric allows us to measure the TBRT current directly.

6.2.1 Evidence of resonant tunnelling

Figure 6.7 shows the results of applying a voltage across the gate stack and measuring current (CG-S) at 300 K. The failure of the gate dielectric allows access to measurements of carrier transport through the TBRT device region. The resonant tunnelling peaks are clearly observed around 1.4 V, which are followed by the sudden decline in current (negative differential resistance) associated with resonant tunnelling. The failure of the gate dielectric reduces the thickness of the structure, thus shifting the peaks to a lower voltage than one would expect for a full memory gate stack. Later, in the next chapter, the tunnelling region is investigated independently from the rest of the gate stack, where simulated TBRT currents emerge at ±1.0-1.6 V. Thus, it is evident that the sudden peak in current observed in positive and negative direction coincides with resonant tunnelling through the InAs/AlSb TBRT junction in the program and erase tunnelling directions respectively. There is a significant current observed prior to the emergence of resonant tunnelling peaks. This is likely to be a result of hole currents due to the small (0.1 eV) valence band offset of InAs/AlSb heterostructure, allowing holes to flow through the gate stack in a similar fashion to initial prototypes without ALD gate dielectrics [216].

6.3 Sample C

Sample C is a memory chip with an intact ALD-Al_2O_3 gate dielectric of 15 nm, processed from the same GaSb/Si growth as sample B. The differences in the fabrication process are summarised in

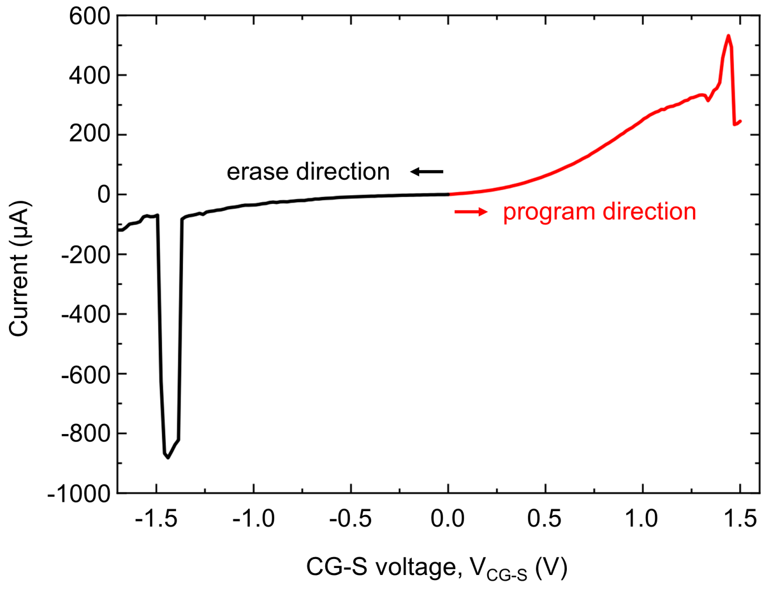


Figure 6.7: I-V sweep of CG-S voltage with measured CG-S current in the program cycle tunnelling direction (red) and erase cycle tunnelling direction (black).

Table 5.2. Titanium-gold contacts were deposited by thermal evaporation for all device terminals. Buffered oxide etchant (10:1) was used to selectively etch AlSb over InAs when etching through the TBRT to access the channel layer. The channel is n-InAs and is therefore NORMALLY-ON and the Ti adhesion layer for the gold-silica interface is substituted for a 2 nm layer of ALD-Al_2O_3 for improved process control.

6.3.1 Channel measurements

The changes to the fabrication process to accommodate the new memory design renders the TLM bars ineffectual, as they receive a double HF exposure which causes significant damage to the layer. Consequently, the channel must be investigated from the memory devices only, using I-V measurements from the devices of 10, 20 and 50 µm gate lengths, as shown in Fig. 6.8. The contact resistance problem identified in previous samples has been eliminated resulting in an Ohmic channel relation as expected from Ti-Au contacts on an n-doped InAs layer. However, there is a leakage current between S-D to the BG terminal of the device (mA), which is unexpected from the InAs/GaAs/AlSb band offsets. In attempting to identify the correlation of S-D current with gate length, an area dependence or direct spacing dependence does not emerge, demonstrating that the majority of the current is not from direct leakage into the BG layer or the channel layer respectively. The most convincing interpretation of the results is when the current density is calculated from the perimeter of the device mesa [Fig. 6.8**(b)**]. This suggests that the leakage current could be the result of surface currents, resulting from the extremely shallow mesa isolation etch ($<$40 nm) used in this sample.

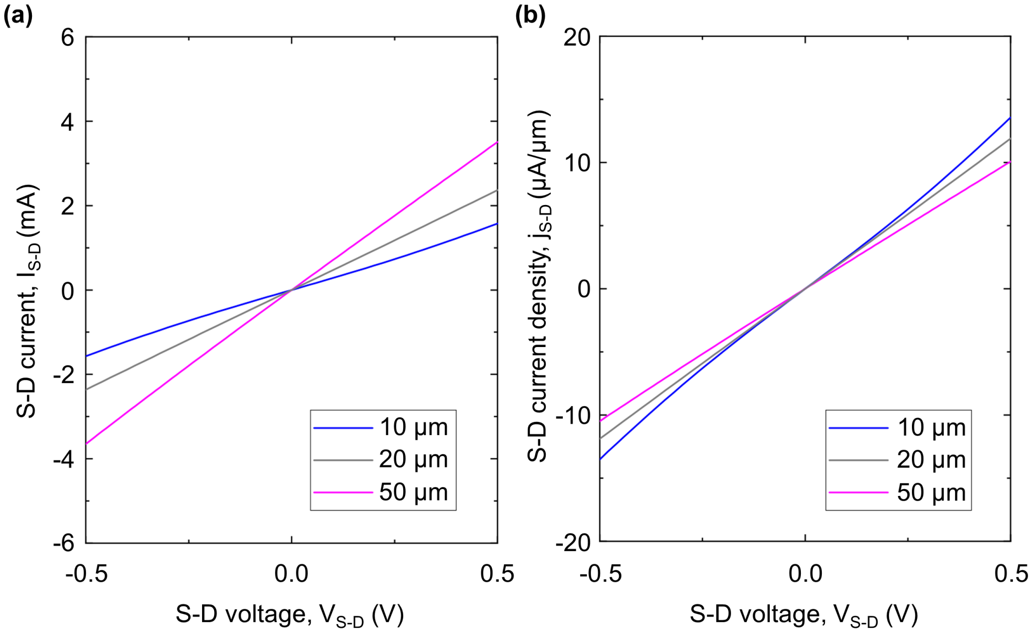


Figure 6.8: S-D I-V sweeps for different feature gate lengths on sample C. **(a)** S-D current measurement for 10, 20 and 50 μm devices. **(b)** S-D current density, calculated as the current per unit length using the perimeter of the device mesa.

6.3.2 FG memory operation

Low-voltage memory switching is demonstrated on single devices fabricated on silicon substrates of 10 and 20 μm devices, where P and E correspond to a charged (0) and discharged (1) FG respectively. The memory operation is shown in Fig. 6.9, where the S-D current is measured at a consistent S-D bias of 0.1 V whilst sweeping through the CG-BG voltage, $V_{\text{CG-BG}}$. As expected, a negative CG-BG bias produces carrier depletion of the n-type channel layer, similar to that observed in an n-type MOSFET. This results in a modest, but observable, decline in current with increasing negative CG-BG bias. When the device is in the P state (0), the negative charges on the FG enhance the negative CG bias such that channel depletion occurs at a lower $V_{\text{CG-BG}}$. The shift in CG-BG voltage is known as the threshold voltage shift, ΔV_{T}, and is the basis for readout measurement of a FG memory, as described in 2.2.1.

The difference between the P and E states of the device shown here gives a ΔV_{T} of around 0.5 V, centred at a CG-BG voltage of approximately -1.25 V (Fig. 6.9). Here, the P and E cycles are performed within the CG-BG sweep by extending the sweep to ±2.5 V and performing the sweep starting from +2.5 V and then -2.5 V to perform P and E cycles respectively and form the memory hysteresis. This technique allows the memory state to be read immediately after it has been set within the sweep. As such, the ΔV_T measured here represents the initial memory state with maximum window and is not an indication of the non-volatility of the memory, which will be discussed later in 6.3.3. Note that the joining of the states at positive bias ($>$ 0.8 V) does not indicate P/E operations: the joining of the states is due to saturation of the carrier population in the channel. The gate-response investigation of the device demonstrates a FG logic state that can be formed at low voltage and read non-destructively.

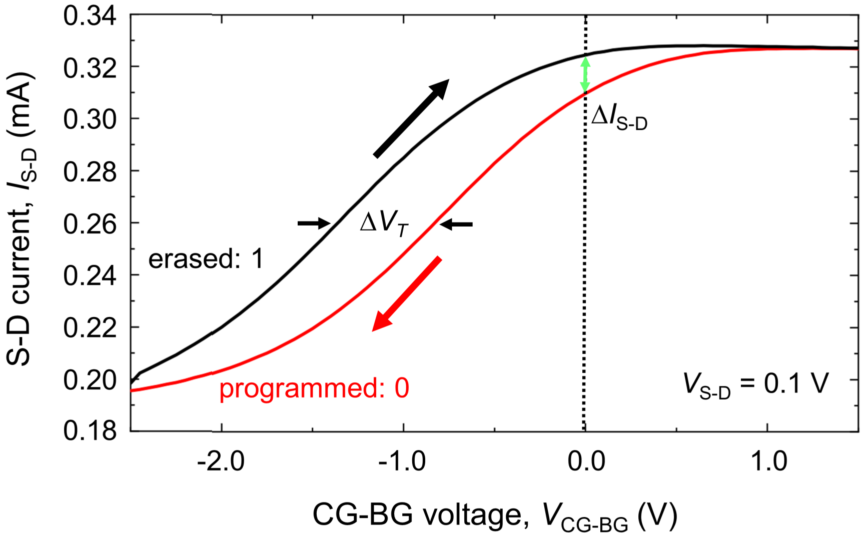


Figure 6.9: S-D current measurement at constant S-D voltage, V_{S-D} = 0.1 V, for a CG-BG voltage sweep from +2.5 V to -2.5 V (red line) and -2.5 V to +2.5 V (black line), representing P and E cycles respectively and forming a FG-memory hysteresis.

C-V measurements

FG-memory operation is further demonstrated by sweeping the CG-BG voltage and measuring the gate capacitance (C-V). The results of a 50 µm device are presented in Fig. 6.10, where all measurements are performed at 1 MHz. Here, the memory state is switched using short (10 ms) pulses of ±2.5 V. A wait period of 10 minutes was adhered to in between C-V sweeps to ensure measurement of a non-volatile state. The C-V was performed after 10 P and E cycles and show a highly repeatable shift in the voltage relation of around 0.25 V. The reduction in in the C-V measurements (Fig. 6.10) compared to I-V measurement (Fig. 6.9) is due to the retention wait time between the P/E pulse and the C-V sweep. Indeed, the memory window has significant decay over this time period, which is observed directly in the next section.

6.3.3 Retention

The retention of the memory logic was investigated on a 20 µm gate length device by repeated measurement of S-D current, I_{S-D}, at a S-D voltage of 0.2 V, but in the absence of a CG-BG bias. The simplicity of this readout scheme is made possible by the n-type, NORMALLY-ON InAs channel design. Note that the removal of a CG-BG bias lessens the size of the memory window in a current

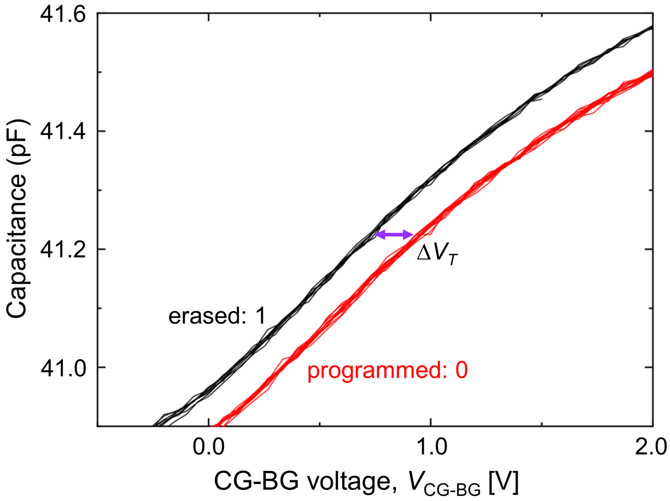


Figure 6.10: C-V measurements after P and E cycles for a 50 µm gate length device on sample C at 1 MHz. Data is magnified to observed the memory window, where the full sweep is $\pm$2.5 V.

measurement, $I_{\text{S-D}}$, as depicted by the green arrow in Fig. 6.9. The readout scheme employed here is favoured for experimental simplicity.

Memory retention was confirmed for >24 hours[4] using $> 10^6$ readout operations for both programmed and erased states, where $\pm$2.5 V pulses of 10 ms duration were used to switch logic. The results are shown in Fig. 6.11**(a)**, where an initial decay in the $\Delta I_{\text{S-D}}$ window is observed, before it fully plateaus at around 22 µA after around 10 hours. Assuming the transconductance of the channel gate response (CG-BG) is consistent as in Fig. 6.9, this corresponds to a non-volatile memory window with $\Delta V_T \sim$ 340 mV. The memory retention was further investigated by plotting $\Delta I_{\text{S-D}}$[5] against time on a log-scale, shown in Fig. 6.11**(b)**. The plateauing of the memory state makes estimation of retention time difficult, as a fit to the data extends to infinity (dashed line). A linear fit (solid line) is made to the gradient of the decay prior to the plateau of the window, and is extrapolated to $\Delta I_{\text{S-D}} = 0$, *i.e.* when the memory window closes. This provides an extremely conservative lower limit of the memory's retention capabilities, as the stabilisation of the memory window is ignored in this fit. Nevertheless, this predicts a memory retention of 10^7 hours, which is more than 1000 years.

The memory decay observed here was not present in sample A (6.1). However, a very similar, but more prominent, state decay was observed in initial memory prototypes on GaAs substrate [216]. The return of the partial state decay for the first devices fabricated on Si substrates suggests some correlation with material quality, as the defect density for the GaAs substrate material used in sample A is a significant improvement compared to the initial prototypes of [216], as summarised in Section 5.1. The most likely mechanism for state decay is charge trapping at

[4] Limited only by the length of the experiment.

[5] $I_{\text{S-D}}$ for state 1 $-$ $I_{\text{S-D}}$ for state 0.

defect sites on the semiconductor heterojunction interfaces. Thus, we predict the elimination of state decay with continued development of epitaxy on Si substrate, as previously achieved with GaAs-substrate growths [233].

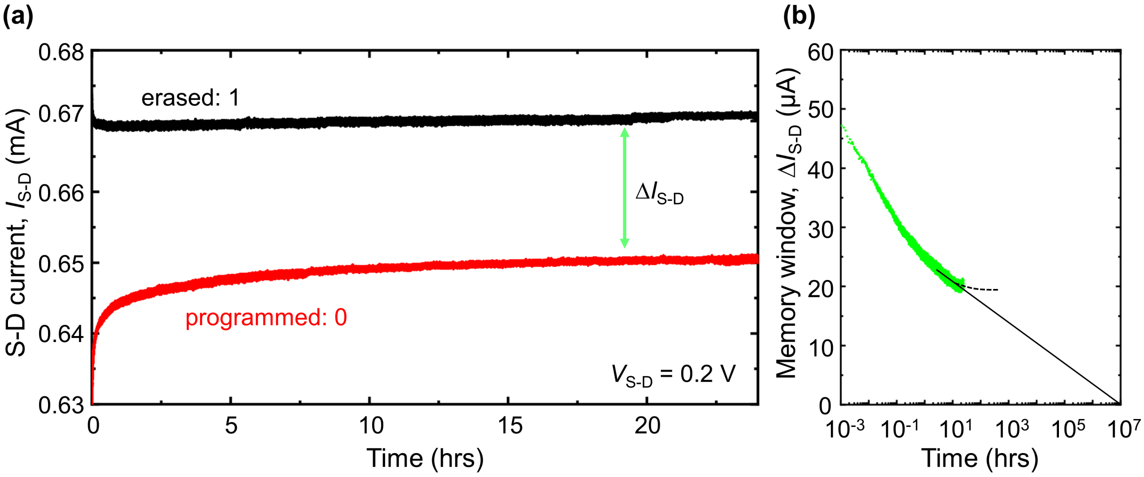


Figure 6.11: Non-volatility testing for sample C. **(a)** Retention data for a 20-μm-gate-length cell. P and E cycles consisted of 10 ms duration pulses at +2.5 V and -2.5 V respectively. Readout is performed at a S-D bias of 0.2 V. **(b)** S-D current difference ($\Delta I_{\text{S-D}}$) for the >24 hour retention plotted on a log scale. Results are discussed in detail in the text.

6.3.4 Endurance

Endurance testing was carried out by P-read-E-read cycling on a fresh 20 μm device using pulses of +2.1 V and -2.55 V (CG-D) for P and E respectively, with S-D current measurements collected at $V_{\text{S-D}}$ = 0.2 V, again in the absence of CG-BG bias. When performing repeated cycling on the memory devices it was apparent that the state was not switching back to its original position when the same P/E bias was used. This is somewhat expected due to the asymmetric construction of the TBRT region. Careful tuning of the P and E bias produced an even P/E process with +2.1 V and -2.55 V pulses respectively, both of 5 ms duration. The lower-voltage requirement of the P cycle corroborates the simulation results for tunnelling current, where the P cycle current peak sits around 0.3 V behind the erase peak (4.5.1).

The memory cell underwent 10^6 P-read-E-read cycles without degradation and a stable memory window, as shown in Fig. 6.12. The cell had zero full-cycle failures and <50 partial cycles. Importantly, the nature of the $I_{\text{S-D}}$ values is highly reproducible. The drift seen in initial prototypes [216] is eliminated and the fluctuation in S-D current seen in sample A is eliminated entirely. The latter is attributed improved processing of the channel etch procedure, as discussed in the previous chapter. The drift is eliminated by tuning the P/E voltages in order to prevent over-programming, which is realised from understanding the asymmetry of the resonant tunnelling process. Indeed, the memory states quickly stabilise once the correct voltages are identified. The remaining drift across the thousands of cycles will become inconsequential when the FG-memory is implemented in a NORMALLY-OFF readout scheme.

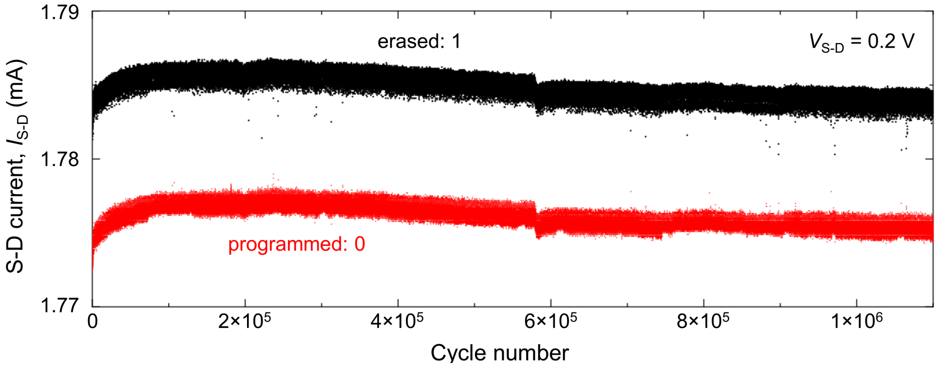


Figure 6.12: P-read-E-read endurance data for a 20 µm device with 5 ms pulses demonstrating clear 0/1 contrast for over 10^6 cycles.

The endurance testing on the same device is extended to 10^7 cycles using a slightly modified methodology. Read operations cause an SMU delay which slows down the cycling speed for endurance measurements. In order to substantially speed up cycling, P/E cycles are applied without a read operation in between as a continuous pulse train. A section of this pulse train is shown in Fig. 6.13**(b)**. Firstly, an initial P-read-E-read test of 1000 cycles is performed in order to establish memory operation, as shown in Fig. 6.13**(a)** in the plot furthest left. The pulse duration for the cycles here are 1 ms. A P/E pulse train of 1 ms pulse durations is then applied continuously for 2×10^6 cycles. Next, 1000 P-read-E-read operations are performed with 10 ms pulses[6] in order to determine if any damage or degradation to the memory cell is apparent. This process is repeated until a total of 10^7 P/E cycles have been performed, shown from left to right in Fig. 6.13**(a)**. The results demonstrate that there is no degradation to memory operation ($\Delta I_{\text{S-D}}$ window) throughout the 10^7 cycles. Thus, this represents an endurance capability that is a minimum of two to three orders of magnitude improvement over flash [75]. However, it is possible that any failure mechanism present is hidden by the large feature size of the device, so endurance tests on scaled devices is required to fully investigate the endurance capability at useful bit densities. Nevertheless, a flash-like SILC failure mechanism causes memory degradation even in large feature size devices, which are not observed in this technology.

In all testing of sample C, P/E switching was performed with pulse durations of 1-10 ms. This is twice as long as pulses required for switching in sample A on GaAs substrate. In both cases, the devices operate at remarkably high speed for their large feature size. Indeed, assuming capacitative scaling the switching speed outperforms DRAM at modern-day feature sizes. The small loss in performance again suggests a modest reduction in material quality on Si compared to the more mature GaAs method used in sample A. Note that the devices on Si still represent

[6] Larger pulses allow easier detection of device failure.

a 1000x speed improvement compared to initial GaAs substrate prototypes in [216]. The most likely reason for defect-related performance losses in charge trapping at defect sites, as this can contribute to *RC* time constant [269] and can cause screening of the applied voltage.

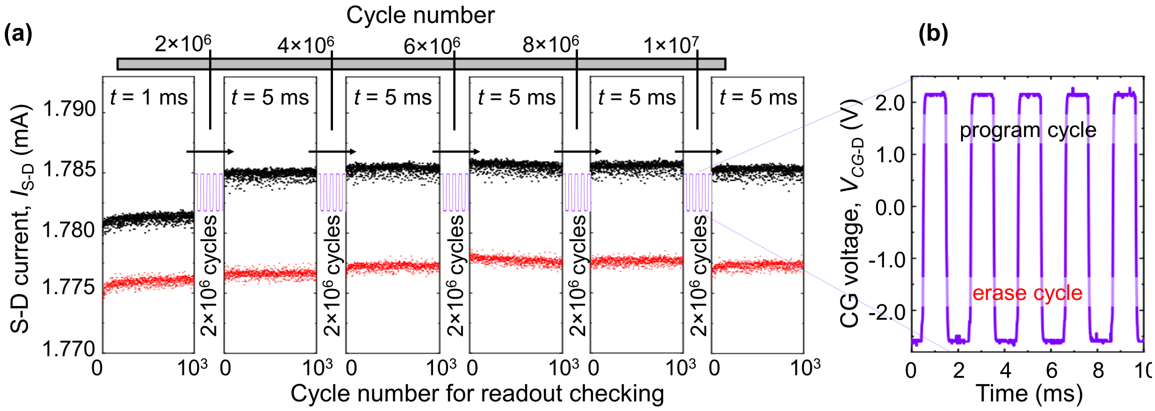


Figure 6.13: Extended endurance of a 20 µm device. **(a))** 10^7 cycles are demonstrated by repeating 2×10^6 no-read P/E cycles five times, where in between each group of cycles 1000 P-read-E-read cycles are performed to confirm memory operation. **(b)** Oscilloscope trace of a section of the applied CG bias pulse train for P/E during extended cycling.

6.3.5 Discussion

Sample C demonstrates ULTRA**RAM** ™ memory on Si substrates for the first time. Non-volatility is confirmed and is predicted to last for at least 1000 years, as electrons are confined behind the 2.1 eV energy barriers of the InAs/AlSb heterojunction. Endurance is found to be at least two orders of magnitude higher than flash memory, owing to the low-voltage, ultra-low-energy (per unit area) operation of the devices. CG-BG response confirms the presence of a threshold voltage window (ΔV_T) which is an essential characteristic of a FG memory. Unlike sample A, the memory logic can be switched with pulses across CG and D terminals due to vastly improved contact resistance. This allows low-voltage cycling without passing any current through the device ($< 10^{-11}$ A). The main areas of improvement have been identified: the initial state decay and slightly weaker cycling speed are attributed to material quality such that developments to III-V epitaxy on Si substrates should provide performance upgrades. The simplicity of the channel design hinders the available 0/1 contrast as the gate response is weak. A channel design with a better ON/OFF ratio would maximise S-D current contrast between 0/1 by many orders of magnitude using the threshold shift between states. Indeed, efforts hereafter focus on the implementing the channel design presented in Section 4.5, which should produce the desired characteristics for a high-contrast FG-memory using III-V layers without the need for lateral doping.

6.4 Sample D

The final ULTRA**RAM** ™ memory chip presented in this thesis is an attempt to utilise the NORMALLY-OFF channel design presented previously in order to improve contrast in the readout measurement of the logic states (*v2.3*). Sample D uses a 5 nm undoped InAs layer as the channel, resulting in a QW ground state energy that is positioned above the GaSb VB to produce an OFF state. An ultrathin 1.2 nm layer of AlSb is grown in between the InAs channel and underlying GaSb as this is found to improve etch quality when using a selective channel etch. The grown wafer includes four sets of AlSb/GaSb dislocation filters in order to improve material quality on the Si substrate. An ALD-HfO_2 gate dielectric of similar thickness to previous samples is used in an effort to widen the threshold window, ΔV_T. The threshold window should improve due to the increased capacitance from the larger dielectric constant, which results in a greater ΔV_T, as outlined in 2.2.1.

6.4.1 Channel current

In implementing a NORMALLY-OFF channel, naturally the first measurements performed on the freshly fabricated devices are I-V measurements from S to D, which should yield a very low current. Unfortunately, this is not the case: devices of sample D have a low resistance channel with an Ohmic response [Fig. 6.14**(a)**]. In order to understand this unexpected result, every memory device on the chip is measured for channel conductivity of 10, 20 and 50 µm gate lengths. The calculated resistance of each device is represented by a point in Fig. 6.14**(b)**. Although some correlation with gate length is observed, it is weak with large overlap between feature sizes such that a dependence on device area, width or channel length cannot be extracted. Moreover, the S-D resistance of the devices is, in some cases, less than half the resistance of similar devices on sample C [Fig. 6.14**(c)**]. As the channel of sample C is thicker and intentionally doped, it is unlikely that the low-resistance S-D connection of sample D is due to carrier transport through the channel layer. Note that channel to BG resistance is also Ohmic, and the hafnia gate dielectric is intact on all devices[7]. The channel conductivity was also measured at low temperature down to 77 K with minimal difference from the 300 K measurements presented here. Possible origins of the unwanted S-D conduction could be surface currents from the extremely shallow mesa etch or conduction through the n-InAs BG layer to channel through etch pits or defect sites (stacking faults). The attempt to implement the previously-simulated channel physics was unsuccessful and faces technical challenges that could be overcome by first understanding the origin of the leakage current, which will be discussed later.

[7] Gate leakage does not contribute to S-D current.

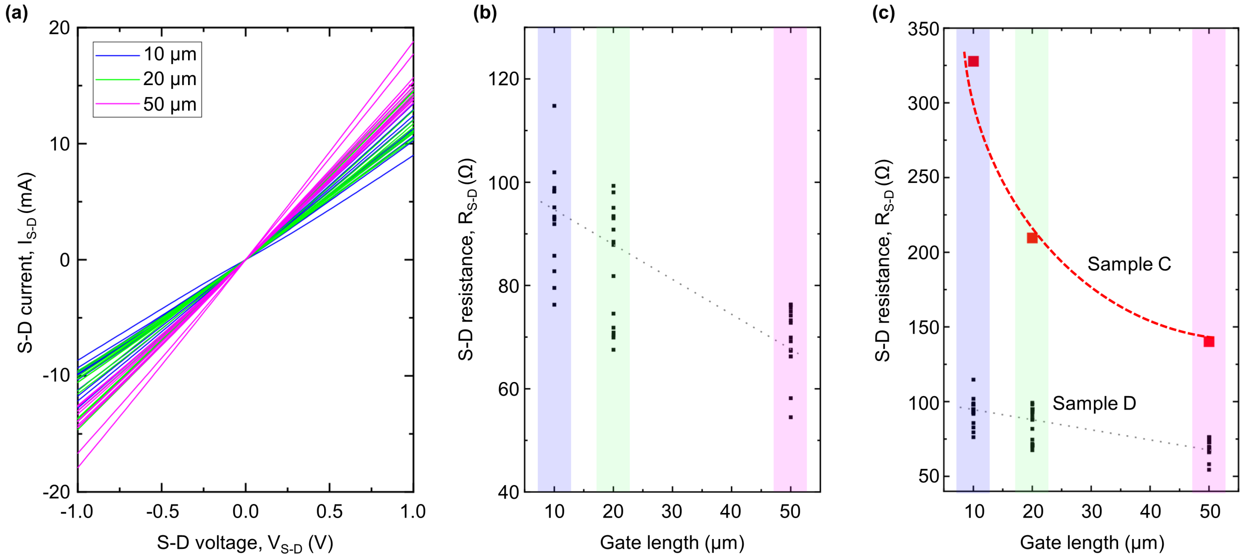


Figure 6.14: S-D current measurements (300 K) of single devices on sample D. **(a)** I-V sweeps of many devices of various feature sizes. **(b)** S-D resistance of each device at different gate lengths on the chip. **(c)** S-D resistance with sample C measurements at similar gate length (red squares) presented alongside for comparison. The lines in **(b)** and **(c)** are guides to the eye.

6.4.2 Memory operation

Despite the extremely large OFF state current, the devices exhibit a gate response and memory hysteresis as a NORMALLY-ON memory in a similar fashion to the previous sample. The memory hysteresis for a ±2.5 V CG-BG bias sweep in shown in Fig. 6.15, where the $I_{\text{S-D}}$ during the sweep is measured at 0.2 V S-D bias. The immediate threshold voltage shift is increased compared to the previous sample as a consequence of the higher-k HfO_2 gate dielectric, which gives a threshold shift exceeding 1 V. This suggests that the memory window can be engineered through the choice of gate dielectric, where the gate dielectric material and thickness could be chosen to suit the desired application if the initial volatile state decay can be successfully eliminated as previously observed in Sample A.

Pulse duration dependence

Although a large memory window is observed in the hysteresis sweeps, the memory state decays over time to stabilise at a much smaller 1/0 contrast ($\Delta I_{\text{S-D}}$). This is again very similar to the previous sample on Si substrate (sample C), and indicates that the DFs used in this growth do not prevent the initial loss of contrast, despite the slight improvement in defect density. In order to investigate the decay in the memory window, 60 minute retention tests were performed on a 20 µm device with 0.1 V S-D bias current measurement in the absence of applied CG bias. P/E states

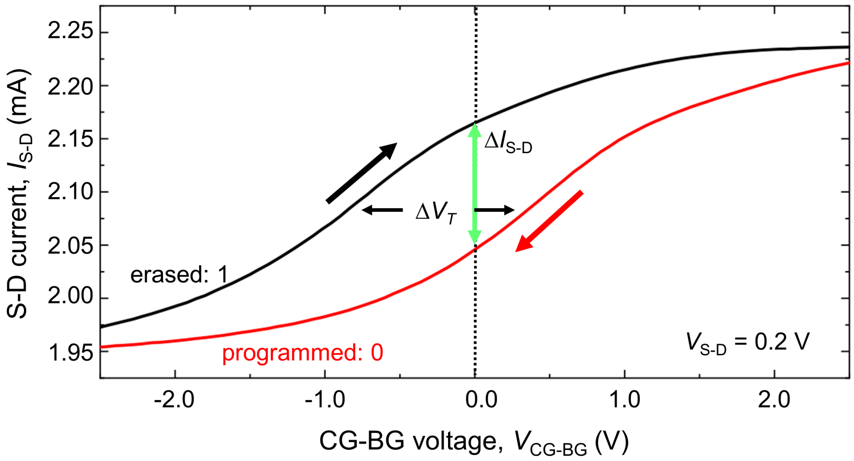


Figure 6.15: S-D current measurement at constant S-D voltage, $V_{\text{S-D}}$ = 0.2 V, for a CG-BG voltage sweep from +2.5 V to -2.5 V (red line) and -2.5 V to +2.5 V (black line), representing P and E cycles respectively and forming a FG-memory hysteresis.

were set using ±2.5 V pulses of increasing durations ranging from 1 ms to 1 s. The results are presented in Fig. 6.16: the memory window plateaus within 60 minutes for all pulse durations and a non-volatile memory logic is observed, indicating non-volatile state-switching at just 1 ms. The time taken to stabilise the memory window is significantly faster than the previous sample. The size of the window is dependent on the pulse duration, where a longer pulse produces a larger non-volatile window. The amount of decay in $\Delta I_{\text{S-D}}$ in the first 10 minutes is similar in all pulse durations used to set the P/E states. The physical process of the gradual state loss is still unknown, as is the reason for the order of magnitude difference in decay time constants between sample C and D. Interestingly, the observation of robust memory windows that are dependent on pulse duration is an encouraging result toward multi-state storage capability. Although, a high-contrast readout operation with the state decay eliminated[8] would be required for multi-state storage capacity.

P/E voltage dependence

Figure 6.17 demonstrates the dependence of the memory window size ($\Delta I_{\text{S-D}}$) with P/E voltages. In this investigation, P-read-E-read cycles are continuously carried out on the memory cell with 10 ms P/E pulse durations and $V_{\text{S-D}} = 0.1$ V readout operations throughout. Here, the P and E voltages are of similar magnitude for simplicity. The experiment begins with ±1.0 V P/E voltages, with many cycles to confirm reliable switching before the next voltage is tested. This is carried out in 0.2 V increments from ±1.0 to ±3.0 V, as labelled in Fig. 6.17. At P/E voltages $\leq$1.6 V, the memory window observed here decays within a few minutes. However, a non-volatile memory state can be obtained with pulses of 1.8 V or larger. This suggests that the initial memory window is made up of two parts: The first is volatile, with a few tens of minutes lifetime, which is likely

[8] As achieved with sample A on GaAs substrate.

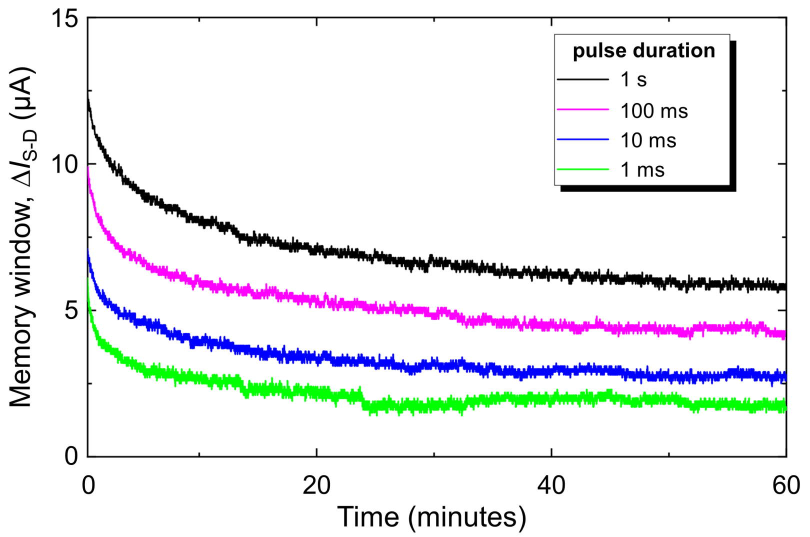


Figure 6.16: Memory window retention over a 60 minute period for P/E pulses of ±2.5 V with 1 ms to 1 s duration. $\Delta I_{\text{S-D}}$ is calculated from the difference in S-D current between P/E states for the similar elapsed time after the P/E pulse at a 0.1 V S-D bias.

to be charge trapping at material interfaces and can therefore be observed at voltages below the TBRT current peaks. The second is the non-volatile state formed from resonant tunnelling into and out of the FG, which occurs when the magnitude of the voltage pulse is large enough for the tunnelling mechanism to occur. There is a clear correlation demonstrated whereby increasing the P/E pulse magnitude increases the size of the memory window. At voltages exceeding 2.6 V, the memory window trends downwards, indicating that over-programming is taking place when equal magnitudes are being used. This can be alleviated by the use of different voltage magnitudes for P and E pulses, as detailed in 6.3.4. The data presented here indicates that the memory window can be made smaller or larger by altering the P/E voltages in a way that is reproducible over 100's of cycles, which could allow multi-state logic by use of different voltage pulses if the technology can be developed into a high-contrast memory with improved $\Delta I_{\text{S-D}}$ stability.

6.4.3 Endurance

Endurance testing was carried out with P-read-E-read cycling on the device using pulses of ±2.6 V of 10 ms duration for P/E cycles, and a S-D voltage of 0.1 V for readout measurement in the absence of CG-BG bias. At the beginning of the test, the device is consistently over-programmed due to the asymmetry of the tunnelling mechanism when using a similar bias for P and E cycling. However, the size and position of the memory eventually stabilises, as shown in Fig. 6.18**(a)**. It is demonstrated that the over-programming of the cell also leads to under-erasing, *i.e.* the entire memory window moves towards a lower current for both P and E states. Before stabilisation, the under-erasing of the cell leaves a small amount of FG-charge behind, which partially screens the

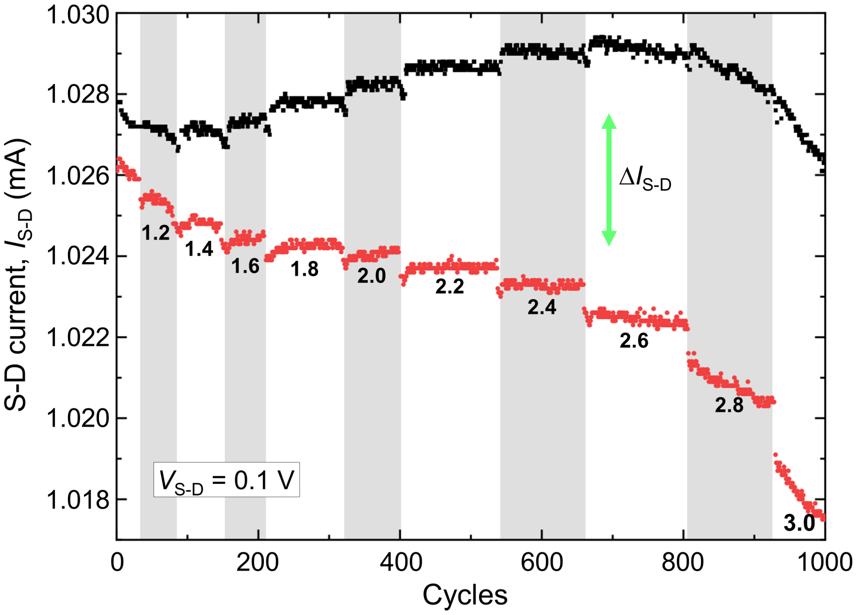


Figure 6.17: S-D current measurements of P-read-E-read cycling for the programmed state (red) and erased state (black) with increasing P/E voltage magnitudes starting from ±1.0 V and increasing in 0.2 V increments.

positive voltage used for the program cycle and partially enhances the negative voltage of the next erase cycle. Eventually, this process symmetrises the P and E cycle, albeit at a slightly smaller memory window size ($\Delta I_{\text{S-D}}$) as plotted in Fig. 6.18**(b)**.

Despite the channel-related issues leading to the small 1/0 contrast, low-voltage logic switching is maintained for over 1.7×10^6 cycles without a single cycle failure. There is a slight dip in the memory window, which occurs four times in the plot shown in Fig. 6.18**(a)**. These are separated by precisely 24 hours of experiment duration. As a reproducible pattern of measurement fluctuation occurs every 24 hours, it is highly likely that the small fluctuations observed here are due to diurnal temperature changes. Importantly, these external factors do not effect the memory window itself [Fig. 6.18**(b)**].

6.4.4 Post-endurance retention

After endurance cycling was finished on the device discussed in the previous section, a retention test was completed to test if endurance cycling degrades the non-volatile capability of the cell. The results of the test on the same device are presented in Fig. 6.19, where S-D current data is collected at a S-D voltage of 0.1 V in the absence of CG bias, and P and E states are set with 10 ms pulses of +2.5 V and -2.5 V respectively. The previous endurance cycling appears to have no adverse effect on the cell's ability to retain a memory window, with the readout current quickly plateauing and remaining stable for $>$10 hours with $> 10^5$ readout measurements.

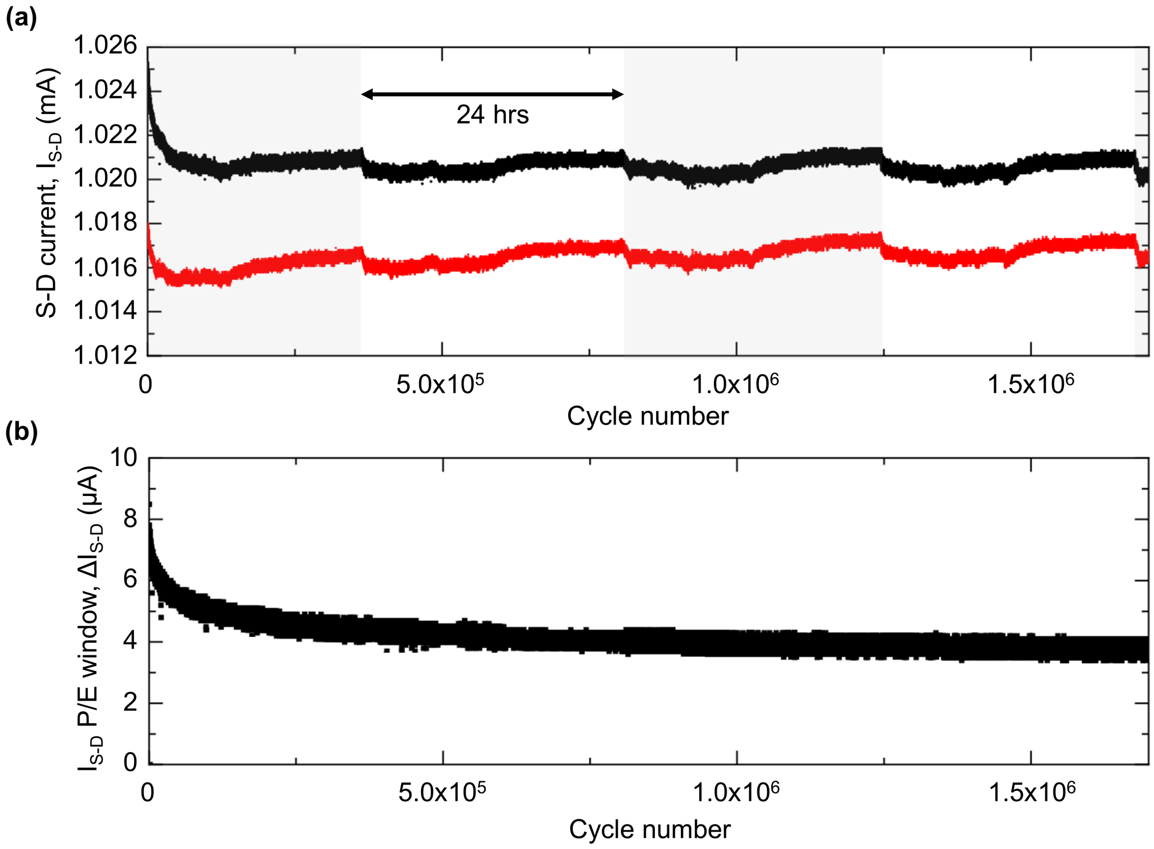


Figure 6.18: P-read-E-read endurance data for a 20 µm device with 10 ms pulses on sample D. **(a)** Readout measurements of S-D current at $V_{\text{S-D}} = 0.1$ V showing reliable state contrast for over 10^6 cycles. **(b)** S-D current difference calculated by subtracting erase and program current from consecutive cycles.

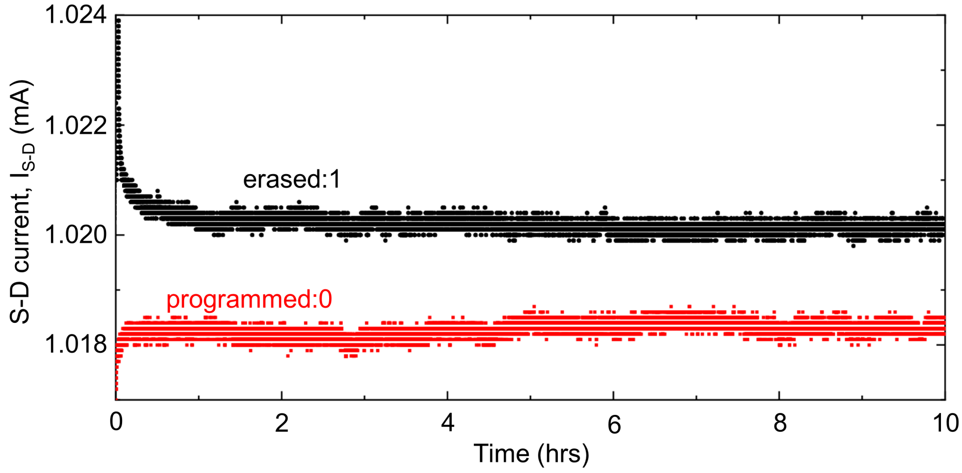


Figure 6.19: Retention data for a 20 µm gate length device on sample D after $> 10^6$ prior P/E cycles on the cell. P/E is performed with 10 ms pulses of +2.5 V and -2.5 V respectively between CG and D,

6.4.5 Threshold voltage window

The previous demonstration of a large threshold voltage window (ΔV_T) sweeps the gate bias from ±2.5 V such that the P/E cycle is contained within the measurement (6.4.2). However, this results in the logic state being measured quickly after it has been set, meaning the threshold voltage calculated from the plot includes the volatile, decaying part of the memory window. In order to measure the non-volatile part of ΔV_T, the gate response is measured after P (10 ms, +2.5 V) and E (10 ms, -2.5 V) cycles in a small CG voltage range (-0.5 V to 0.5 V) with the sweep performed in the same direction for both cycles, and the S-D bias at 0.2 V (Fig. 6.20). After the logic state is switched, a retention test is monitored until the decay in the memory state is complete, and a stable current is observed (<2 hours). Then, the small-range CG-BG sweep is performed on the cell whilst in a stable retention state. A non-volatile ΔV_T of $>$130 mV is observed (Fig. 6.20). Although this is small compared to the main hysteresis curve presented previously, it still represents a possibility for multiple orders of magnitude of 1/0 contrast if the readout scheme can be improved.

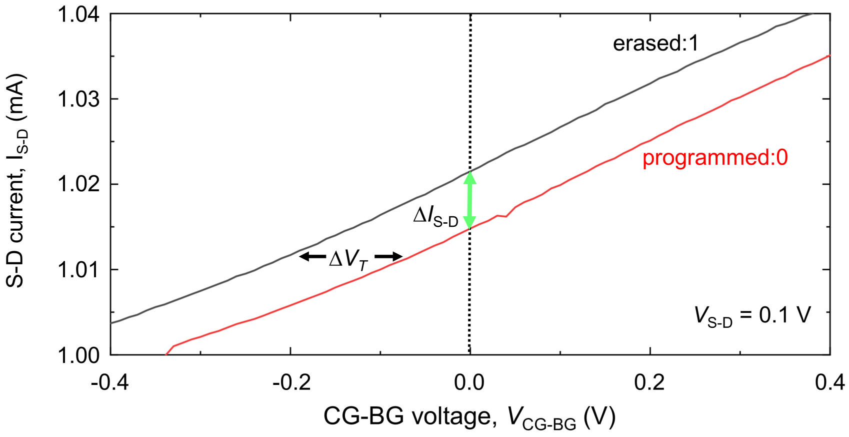


Figure 6.20: CG-BG gate response on a memory cell on sample D after P (10 ms, +2.5 V) and E (10 ms, -2.5 V) pulses. After the pulse the initial state decay was observed to stabilise before starting the sweep in order to measure only the non-volatile memory window.

An interesting observation for this sample is that the non-volatile ΔV_T window is less than half of the previous sample with a very similar layer structure. This is the opposite of what one would expect, as the inclusion of the HfO_2 gate dielectric should improve memory performance: the use of high-k gate dielectrics in conventional FG-memory structures improves the gate coupling ratio which extends the memory window [270, 271]. However, this is based on the assumption that there is a large enough DOS in the FG to accommodate all of the electrons tunnelling into it during the program cycle. As the FG layer is a thin (10 nm) layer of InAs, a QW is formed with only the ground state below the resonant tunnelling energies in the TBRT. A full calculation of the electron 2D DOS is provided in Appendix E, where the maximum FG-charge storage (Q_{FG}) is calculated assuming all states below the resonant tunnelling energies are filled during a P cycle. When this value is used with the ΔV_T relation outlined in 4.6.1 (equation 4.6), an important understanding of the device physics is obtained. The calculations, provided in the Appendix, predict maximum ΔV_T values of 370 mV and 130 mV for samples C and D respectively. Indeed, a quantum interpretation of the FG occupation not only describes the unexpected correlation with gate dielectric capacitance but also produces ΔV_T values that are in close agreement with threshold voltage shift measurements of 340 mV and 120 mV for samples C and D respectively. This results suggest that extension of the non-volatile ΔV_T window requires a larger electron DOS in the FG, which can be obtained by thickening the layer or substituting for an alternative FG material.

6.5 Channel design developments and troubleshooting

As demonstrated by prior attempts to implement a NORMALLY-OFF channel into the technology, unwanted leakage currents between S, D and BG terminals hinders channel performance. To investigate this persistent problem, an ULTRA**RAM**™ wafer was grown with a similar layer design to that of sample D with two important changes:

- The InAs channel thickness is increased to 10 nm, which provides a NORMALLY-ON channel for ease of characterisation.
- The n-InAs BG under the ultra-thin channel is removed, and an intentionally p-doped GaSb layer[9] is added under the deepest DF to act as a potential BG layer that is further away from the channel surface (>500 nm).

This design is presented schematically in Fig. 6.21**(a)** (not to scale), and will be hereby referred to as *v2.4*.

6.5.1 Deep mesa isolation

In order to investigate whether the channel-BG current is a result of insufficient mesa isolation, TLM bars are made on the new wafer where the surface is the InAs channel layer, accessed through the standard selective wet etch procedure discussed previously. After revealing the channel layer on the sample surface, the area surrounding the TLM structure is ICP-etched (5.4) through all remaining III-V layers to the n-type Si substrate. The surfaces are then passivated with 20 nm of ALD-Al_2O_3 before windows are opened with a BOE-dip and Ti-Au contacts are added (shown schematically in Fig. 6.21**(b)**).

The TLM results are displayed in Fig. 6.22**(a)**. The sheet resistance of the undoped 10 nm InAs channel is 163 $\Omega/\square$, and contact resistance was 35 Ω, both calculated from the fit of Fig.

[9] Beryllium doped, $p \sim 2 \times 10^{18}$ cm^{-2}

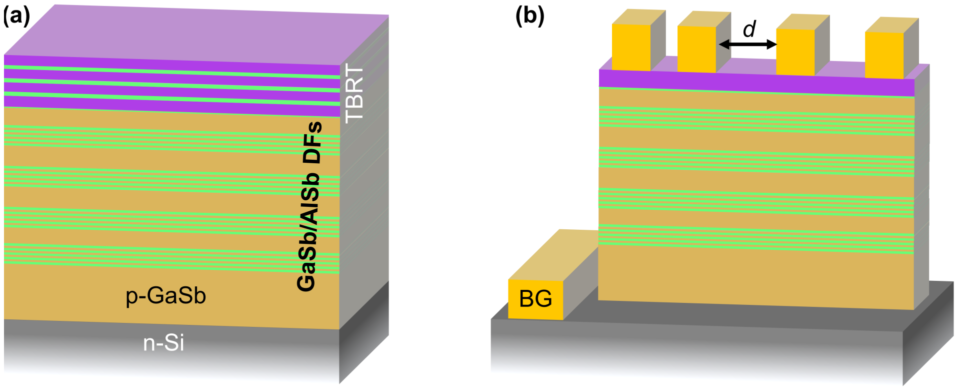


Figure 6.21: Schematic layer design of the growth used to investigate the channel-BG leakage problem (not to scale). **(a)** Full *v2.4* wafer design, with the BG layer changed for p-GaSb positioned below the DFs. **(b)**. TLM bar schematic, with deep mesa isolation and Ti-Au contacts on the channel layer surface.

6.22**(a)**, using the method previously outlined in 6.1.1. The TLM results presented here are the first example of a memory layer TLM measurement exhibiting the expected linear relation with contact separation. This suggests that the ultra-shallow mesa isolation was the cause of the extremely low channel-BG and S-D resistances observed in sample D.

Current-voltage measurements where performed across metal BG terminals (Ti-Au) deposited directly onto the etched n-type Si surface [Fig. 6.21**(b)**]. The n-type silicon BG layer gave a diode response due to the Schottky barrier formed at the silicon-titanium interface [272]. Consequently, measurements of the channel-BG leakage are dominated by the high-resistance contacts and are not be a true leakage measurement. Therefore, channel-BG leakage is measured by probing separate InAs-channel TLM structures where surrounding material has been etched to the Si substrate, as previous measurements have shown that an Ohmic channel contact has been made successfully to this layer. The results are shown in Fig. 6.22**(b)**, with a schematic of the measurement technique provided as an inset. The channel-BG leakage is extremely small, and represents a $> 10^4$ improvement compared to sample D. Moreover, the channel leakage here is reduced by a factor of 1000 compared to the channel resistance on the InAs TLM bar [Fig. 6.22**(a)**], which offers some indication of ON/OFF contrast if the channel can be switched OFF with a gate bias.

6.5.2 Channel etch

The channel-BG leakage current observed in shallow-etched devices could be a result of BG shorting through etch pits as a result of poor wet-etching quality. In previous growths, this is difficult to diagnose as the InAs BG layer is positioned just 28 nm below the channel, making target-layer determination for the etch unfeasible. However, the *v2.4* memory growth used to investigate the

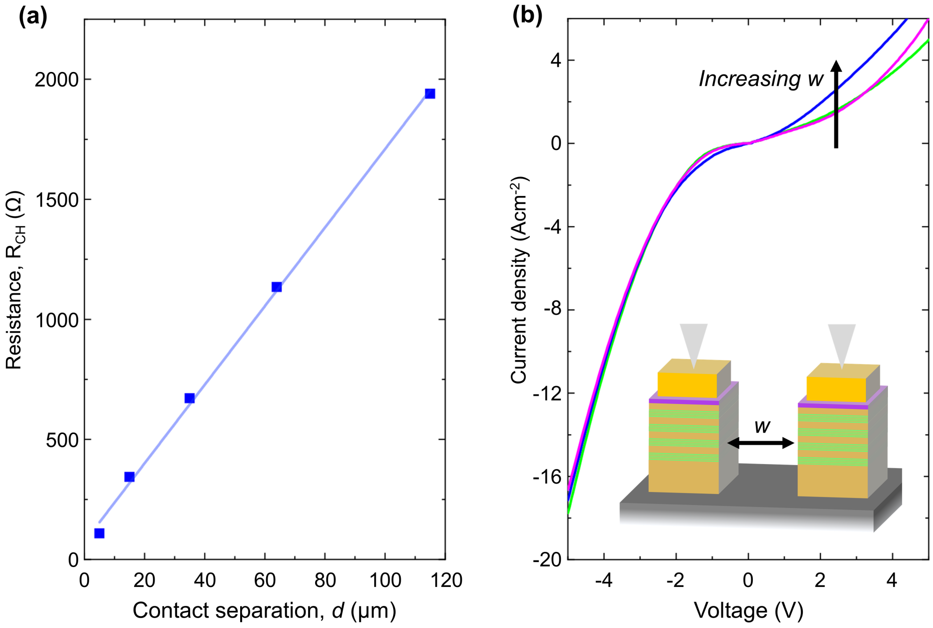


Figure 6.22: Measurement on deep mesa etched TLM bars on wafer *v2.4*. **(a)** Dependence on contact separation on the TLM bar with channel layer resistance, R_{CH} . **(b)** Channel-BG leakage current density measured between TLM structures isolated from one another by the deep mesa etch (presented schematically in the inset).

deep mesa isolation has no such BG layer, as it is now p-GaSb. Indeed, when the TBRT region is etched to reveal the channel, this is the final layer containing In or As elements, allowing for EDX (3.2.1) species mapping to investigate the channel etch quality.

An SEM image of an ICP-etched sample from this growth[10] is presented in Fig. 6.23**(a)**, where the surfaces have also been passivated with a 20 nm ALD-Al_2O_3 layer. The SEM image indicates that the InAs channel surface revealed by selective etching of the tunnelling layers is smooth and uniform, with no etch pits visible[11]. Fig. 6.23**(b)** and **(c)** are EDX maps of the same area for In and As respectively. The absence of data points along the edge of the etched mesa is clearly a result of shadowing from its height. A clear difference in In and As signal is observed between the deep-etched side and preserved mesa side of the sample, where the larger element signal of the left side confirms that the 10 nm InAs channel layer is intact. The small signal on the etched section is likely a result of re-deposition of small amount of reactant products during the plasma etch. The results presented here confirm that a uniform, highly selective channel etch has been developed for ULTRA**RAM**™ on Si substrate, and rules out the possibility that leakage is due to poor process control.

[10] Note that this is the same sample used for the TLM investigation in the previous section.

[11] The presence of etch pits are usually visible even with optical microscopy.

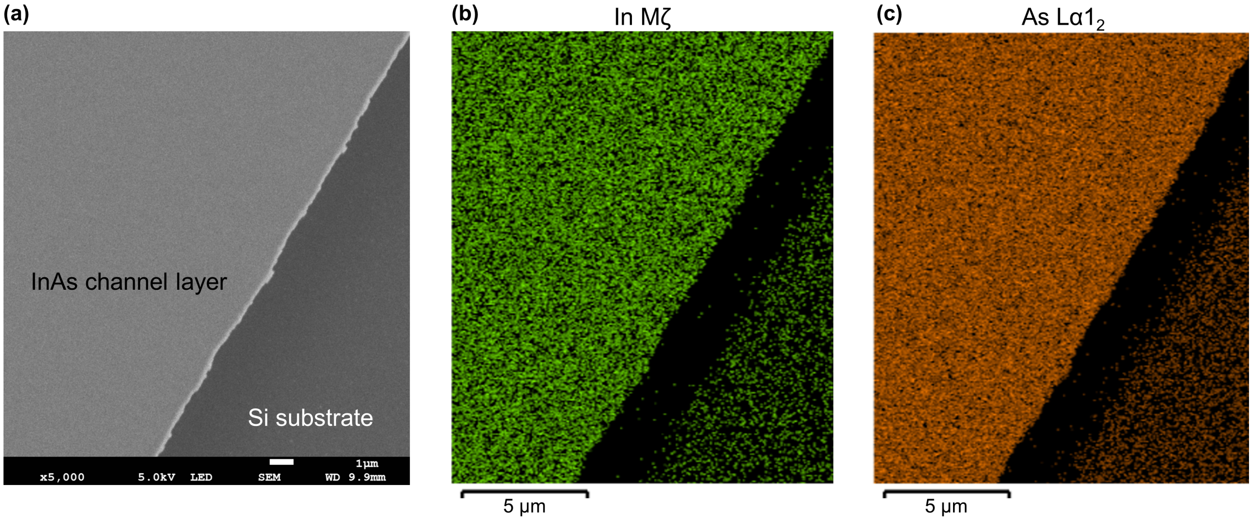


Figure 6.23: SEM/EDX investigation of the selective channel etch procedure using ULTRA**RAM**™ *v2.4* material. **(a)** SEM image of the mesa, where the top of the mesa is the wet-etched InAs channel. **(b)** EDX species mapping of In demonstrating the presence of indium resulting from the channel layer. **(c)** Similar EDX measurements for arsenic. Note that each coloured pixel of the EDX maps represents the verification of the species by the software, *i.e.* the pixel colour is independent of EDX peak intensity once the species is identified at a given point.

6.5.3 Memory window size

6.6 Summary

The principles of the ULTRA**RAM**™ memory concept have been demonstrated for NORMALLY-ON channel configurations at 10, 20 and 50 µm feature size devices on GaAs and Si substrates. Non-volatile memory states are observed which are predicted to last at least 1000 years, and can be switched at low voltages and high speed[12]. Consequently, the switching energy and speed per unit area corroborates the extraordinary memory performance characteristics from previous simulations of the device physics. Moreover, a gate response is observed in which a threshold voltage shift confirms FG-charge storage from both I-V and C-V measurements at room temperature. Memory endurance exceeds 10^7 P/E cycles, which is 2-3 orders of magnitude improvement over flash. However, smaller devices with high-frequency switching capability are required to further investigate the full endurance capability within a reasonable timescale and investigate if there is a scaling relation with memory endurance.

The main challenges of ULTRA**RAM**™'s further development are related to 1/0 readout contrast. The issue results from poor mesa isolation due to the shallow-etched device design. Indeed, deep-etched mesas have a significantly reduced leakage current using the same materials

[12] For the relatively large feature size.

(GaSb/AlSb) for channel-BG isolation, albeit with a greatly increased thickness. With this design change and steady improvements in growth quality, it is predicted that 1/0 contrast will improve drastically, allowing expansion into large ULTRA**RAM**™ memory arrays to commence.

Chapter 7

Simulations of ULTRARAM™ Resonant Tunnelling Heterostructures

This chapter outlines the effects of thickness variations to the ULTRA**RAM**™ InAs/AlSb TBRT heterostructure, which is primary source of its outstanding performance characteristics. The investigation is conducted for performance optimization, and for assessing growth and process tolerances for commercial implementation on 12” Si wafers.

7.1 Alterations to the TBRT

The remarkable memory performance of ULTRA**RAM**™ is predicted by detailed simulations of quantum transport in Chapter 4, and encouraging results have been demonstrated in this thesis on single devices and 2×2 arrays on GaAs and Si substrates at 10-50 µm gate lengths. Our previous theoretical investigations were for a specific layer thickness configuration of the triple-barrier InAs/AlSb tunnelling junction. In practice, growth of these layers with exact monolayer (ML) precision is not straightforward. Indeed, thicknesses could be offset across the entire growth due to imprecise calibration, or vary across the wafer. This is of particular concern for the commercial development of ULTRA**RAM**™, as it is vital that the InAs/AlSb heterostructures currently grown on 3” Si be transferred onto 12” Si substrates in order to be cost-competitive.

In this chapter, we present further NEGF simulations of the ULTRA**RAM**™ resonant tunnelling region, where layer thicknesses are varied to investigate the effect on the performance of the memory. The analysis of these results is compared to the so-called ‘target’ structure, which is the design used for memory performance analysis in Chapter 4. Further details of the simulation method are given in prior sections 3.5.2 and 4.5. The choices of layer alterations presented here are ±1 ML (*i.e.* one lattice constant, 6 Å), and are presented in Table 7.1. It is possible for alterations of 0.5 ML to occur; however, these have a smaller impact on the memory performance.

Table 7.1: Layer thicknesses of the TBRT region (Å). The three AlSb barriers, denoted as B_1, B_2 and B_3, are ordered in the P cycle tunnelling direction (*i.e.* B_1 is adjacent to the channel). InAs layers are positioned between the barriers such that two quantum wells form, denoted as QW_1 and QW_2, where QW_1 is closest to the channel.

	AlSb B_1	**InAs QW_1**	**AlSb B_2**	**InAs QW_2**	**AlSb B_3**
Target	18	30	12	24	18
Equal QWs	18	24	12	24	18
+1 ML QW_1, QW_2	18	36	12	30	18
-1 ML QW_1, QW_2	18	24	12	18	18
+1 ML B_1, B_2, B_3	24	30	18	24	24
-1 ML B_1, B_2, B_3	12	30	6	24	12
+1 ML B_1, B_3	24	30	12	24	24
-1 ML B_1, B_3	12	30	12	24	12
+1 ML B_2	18	30	18	24	18
-1 ML B_2	18	30	6	24	18

Furthermore, the thinnest layer of the tunnelling region has a target thickness of just 2 MLs (12 Å), so an increase of 2 MLs would represent a doubling in thickness, whilst a decrease of 2 MLs removes the layer entirely.

7.2 Non-volatility

Fabricated ULTRA**RAM**™ memories have exhibited encouraging retention characteristics, indicating suitability as a non-volatile memory. Earlier, in Section 4.7.4, the ability of the memory to retain the logic state was analysed by careful consideration of the transmission through the TBRT region at zero applied bias (*i.e.* the electron transparency of the TBRT) which would cause failure of the memory to retain the logic state in the FG. To briefly reiterate the discussion of Section 4.7.4, the intrinsic 300 K (thermal excitation) storage time of electrons in the InAs/AlSb system exceeds the age of the universe [284]. However, to investigate the non-volatility of ULTRA**RAM**™ we must consider the effects of the TBRT structure on the transparency of the barriers. The example of the target structure is repeated in Fig. 7.1**(a)** where there are low-energy transmission peaks corresponding to the energy levels of QW_1 and QW_2 of the TBRT region. Here, the probability of transmission is very low, making it unlikely that these peaks impact on the retention capability of the memory. The observation of stable memory retention in fabricated exceeding 24 hours at 300 K with little or no state decay supports this assertion [233].

Monolayer alterations to the target structure are assessed by repeating the NEGF calcula-

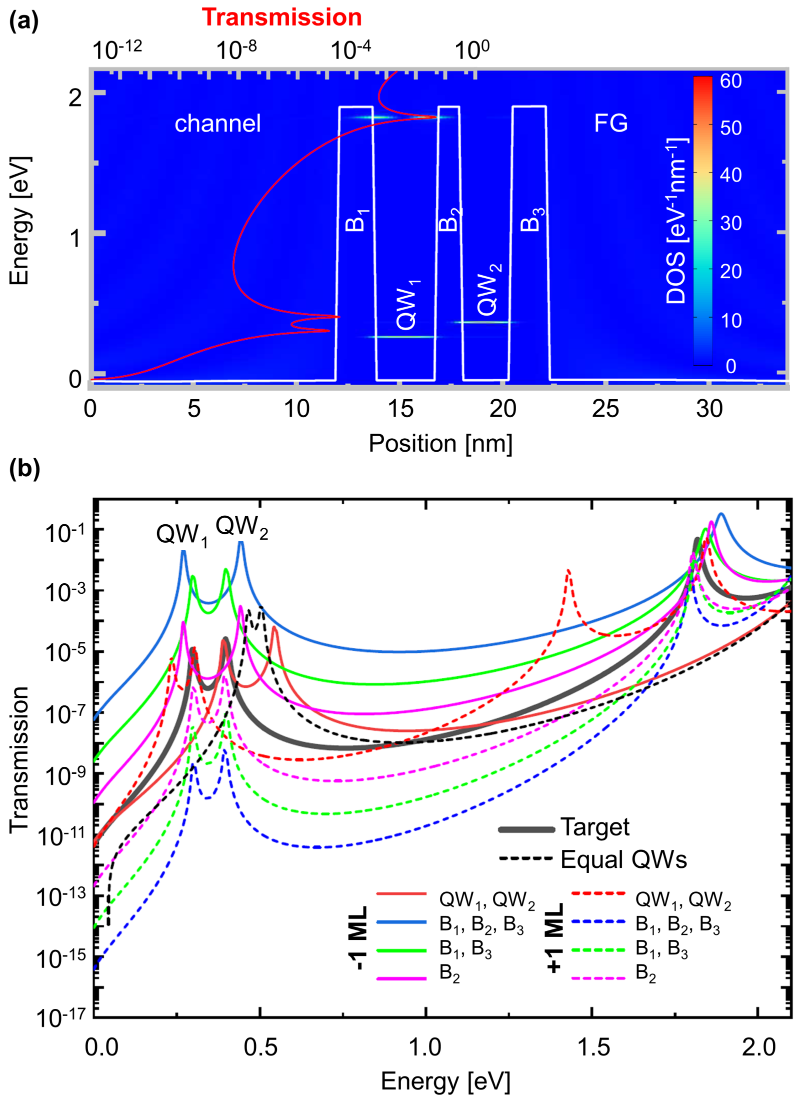


Figure 7.1: NEGF transmission calculations for the ULTRA**RAM**™ TBRT region at 300 K under zero bias. **(a)** Position-resolved, electron energy levels of the QWs for the target heterostructure, where the colour-scale indicates the DOS. The conduction band calculation is shown by the white line. The corresponding transmission function (red line) demonstrates the peak alignments with the confined energy levels in the structure. **(b)** Transmission functions for each layer thickness alteration described in Table 7.1. Results are described in detail in the text. Figure reproduced from [273].

tions under the same conditions for each of the alterations listed in Table 7.1. The results for the transmission function are presented in Fig. 7.1**(b)**. Alterations to the QW thicknesses shifts the energy states of the QWs. A 1-ML reduction in QW thickness shifts the start of the QW_1 transmission peak to a higher energy (solid red line, Fig. 7.1**(b)**) by around 100 meV, with similar transmission, such that it is coincident with the transmission peak associated with QW_2 in the target structure. This is not unexpected as the QW width in both cases is the same. Similarly, the transmission peak associated with QW_2 shifts to a higher energy. The expected result to memory performance would be an improvement in retention compared to the target structure.

The converse argument can be made for increasing QW thicknesses by 1 ML (dotted red line, Fig. 7.1**(b)**), which may decrease retention capability. The transmission peak associated with QW_2 becomes coincident with that of QW_1 in the target structure, in both cases corresponding to a QW thickness of 30 Å, and the peak associated with QW_1 moves to lower energy. The 1-ML increase also shifts the transmission peak associated with the second confined state of QW_1 to lower energy, and introduces a new peak associated with the second confinement energy of QW_2 at similar energy and transmission to the second confined state of QW_1 in the target structure. In summary, if all barriers are at the original target widths, the energies and magnitudes of the transmission peaks are closely associated with given QW widths, and are somewhat independent of each other. If QW thicknesses are equal, the increased overlap interaction between the wells increases transmission by an order of magnitude, as shown in Fig. 7.1**(b)** (black dashed line).

Reducing the thickness of the barriers increases the transmission probabilities. A 1-ML thickness reduction of all barriers increases the transmission probability at the QW_1 ground state energy by about three orders of magnitude to $T = 0.02$, with a potentially detrimental impact on the non-volatility of the memory (solid blue line, 7.1**(b)**). Reducing the central barrier thickness (B_2) results in greater separation between QW_1 and QW_2 peak energies as the energy splitting effect from the interaction of the two QW ground states due to the Pauli exclusion principle is increased [274]. This forces the QW_1 energy lower which could negatively impact retention, however the transmission remains relatively small (solid pink line, 7.1**(b)**).

Unsurprisingly, increasing the barrier thicknesses has the opposite effect, and will result in improved retention capability of ULTRA**RAM**™ if required (dotted blue, green and pink lines, 7.1**(b)**). Successive 1-ML increases in barrier thickness each deliver a reduction in QW peak transmission of about an order of magnitude. There will be a trade-off between the resonant-tunnelling current characteristics (*i.e.* P/E capability) and the charge-blocking properties, so we cannot yet conclude that thickening the barriers is a superior design choice for the technology until the full extent of ULTRA**RAM**™ retention properties have been experimentally investigated. Crucially, we find that the retention of the memory should not be detrimentally impacted by most 1-ML alterations, with the main retention concerns being the cases where thickness is reduced for all barriers. Fortunately, the latter scenario is experimentally relatively unlikely, as it represents a large change in percentage error in layer thickness for layers grown epitaxially in close proximity to each other. This is a promising result regarding the commercial implementation of ULTRA**RAM**™, but demonstrates that process and growth tolerances are of paramount importance.

7.3 Tunnelling current

The TBRT mechanism used to program (*i.e.* add electrons to the FG) for ULTRA**RAM**™ memory devices is demonstrated in Fig. 7.2, where all calculations are again carried out using the nextnano MSB software package [166]. Resonant tunnelling occurs under the condition that the energy of electrons in the channel align with the available energies of the TBRT structure (*i.e.* the QW_1 or QW_2 ground state energies). This is measured as a current density through the barriers, as shown in Fig. 7.2**(d)**, containing three distinct peaks. The DOS plots under applied bias reveal the resonant conditions of the large tunnelling currents [Fig. 7.2**(a)**-**(c)**] as labelled on the current density plot. The low-voltage peak [Fig. 7.2**(a)**] occurs when the electron energies from the contact align with the QW ground states, producing a broad peak. However, the device construction is such that the contact on the channel is spatially separated from the gate stack, therefore resonant tunnelling directly from the contact (*i.e.* ballistic tunnelling that bypasses the InAs CB entirely) is not possible. We conclude that this is not part of the ULTRA**RAM**™ tunnelling mechanism and is an artefact of the simulation construction. Indeed, experimental studies support this assertion [233, 275, 276, 277]. The peaks occurring at higher voltages are the expected resonant tunnelling peaks (Fig 7.2 **(b)** and **(c)**). As the TBRT region is under a large electric field, the sloping of the conduction band forms a triangular quantum potential well at the channel-B_1 interface such that resonant tunnelling through the structure is a 2D-2D process with a concentrated DOS. Here, electrons occupy the triangular quasi-bound state due to inelastic scattering. Without the inclusion of inelastic scattering the device operates in an entirely different way whereby the current-density characteristic is entirely determined by the properties of the lead (contact), rather than by details of the conduction band in the device [278]. Two large tunnelling current density peaks emerge corresponding to alignment with the QW_1 and QW_2 ground states at V_{TBRT} = -1.06 V and V_{TBRT} = -1.28 V, respectively. Tunnelling through the QW_2 state can occur despite the QW_1 state residing at higher energy due to the wave-function overlap between the QWs, which can be seen in the DOS plots of Fig. 7.2**(c)**.

When interpreting a resonant tunnelling current-density plot for the purposes of ULTRA**RAM**™ memory there are important properties to consider. First, there is the magnitude of the current density peaks, which is directly related to the device switching speed [218], as demonstrated in Section 4.7, where larger current peaks improve performance. Second is the voltage (V_{TBRT}) at which the peak occurs; a high voltage will consume more power, whilst a very small voltage is more likely to have logic disturbances under readout biasing. Lastly, is the sharpness of the onset of the peaks; a significant tunnelling current away from the P/E peak voltage will also cause logic disturbances [218]. It can be seen in Fig. 7.2 that the resonant tunnelling current forms sharp peaks at a little over 1 V, indicating that the TBRT structure is highly suitable for the implementation of FG memory, as demonstrated in experimental work outlined in previous chapters. The simulations are repeated under the same conditions for each of the structures listed in Table 7.1 for the program cycle, and are shown in Fig. 7.3. Alterations to the QW thicknesses [Fig. 7.3**(a)**] shift

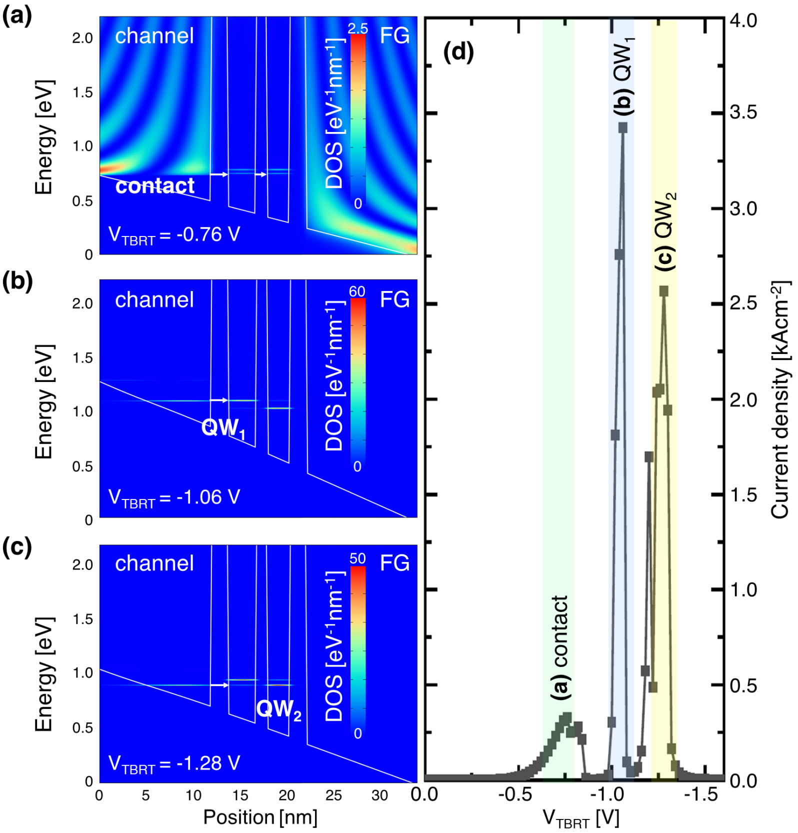


Figure 7.2: Simulations of the target TBRT junction for the memory (300 K) for tunnelling of the program cycle. **(a)** DOS of states for the "leads" (contacts) under bias corresponding to the contact peak of the current density plot. **(b)** DOS (colour scale) plot for the tunnelling voltage of QW_1 current density peak. **(c)** DOS (colour scale) plot for the tunnelling voltage of QW_2 current density peak. **(d)** Current density plot of the TBRT region as a function of applied bias. Labelled peaks correspond to the resonant energy alignments of **(a)**-**(c)**. Figure reproduced from [273].

their ground state energies such that a reduction in thickness increases the required electric field for resonant tunnelling alignments [pink triangles, Fig. 7.3**(a)**]. Moreover, this shift outweighs the increase in electric field caused by thinning the structure. The converse applies when thickening the QWs. The magnitude of the current density peaks is not significantly affected. Thus, changes to the QW thicknesses could serve as a valuable method in tuning the memory for operation at a desired voltage without hindering performance.

Fig. 7.3**(b)** presents the results of thickening barrier layers on the P cycle current density. Increasing barrier thickness reduces the current density by up to two orders of magnitude, which could degrade the speed of the memory. However, a trade-off could be made with the retention capability such that speed is exchanged for more robust logic retention as discussed in the previous section. It is worth noting that this speed reduction may be significant, but it is likely that ULTRA**RAM**™ would still outperform current memory technologies as the tunnelling mechanism is intrinsically fast [218, 233, 279]. Increasing the thickness of the middle barrier [red dots, Fig 7.3**(b)**] greatly reduces the tunnelling current from QW_2 (compared to QW_1) due to the reduction in wave-function overlap between the QWs, with the ground state of QW_1 above the energy QW_2 [similar to Fig. 7.2**(c)**]. Logic disturbances at low voltages within the architecture can be evaluated from the region below the peak, depicted by the orange shading in Fig. 7.3**(b)**, where we ignore tunnelling from the contact for reasons previously discussed (labelled 'off'). Increasing barrier thickness reduces the current density in the off region which could be used to reduce logic disturbance. However, it should be noted here that testing on fabricated devices from the previous chapter retained robust logic states after 10^5 disturbance cycles.

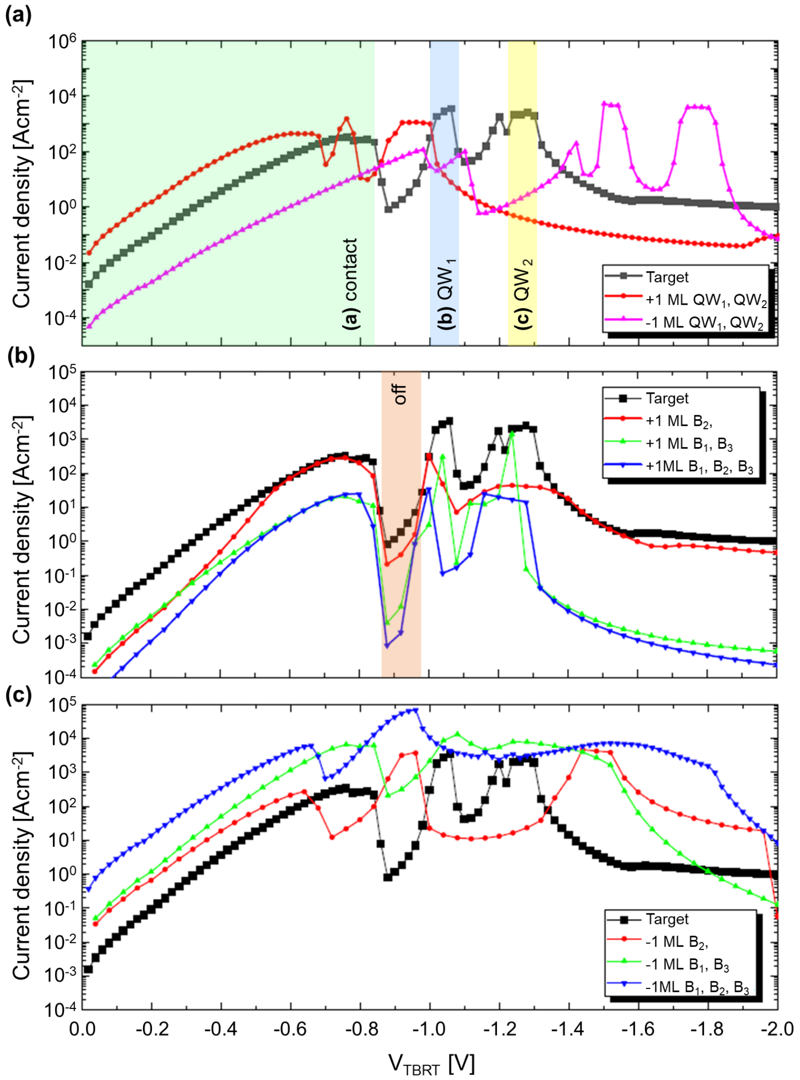


Figure 7.3: Nextnano MSB NEGF simulations (300 K) of TBRT current density for the program cycle for each of the alterations listed in Table I. Results are described in detail in the text. **(a)** 1 ML alterations to the QWs. **(b)** Increased barrier thicknesses. **(c)** Decreased barrier thicknesses. Figure reproduced from [273].

Reducing the barrier thicknesses increases the current densities, as shown in Fig. 7.3**(c)**.

However, the maximum increase from these alterations is just over one order of magnitude, so switching speed benefits must be carefully weighed against other requirements of the memory. Decreasing the thickness of the outer barriers, B_1 and B_3, slightly increases the current-density (Fig 7.3**(c)**, green triangles), but broadens the peaks and produces a more gradual ramp to peak current (with omitted contact tunnelling *i.e.* the in the 'off' region). This could negatively impact logic disturbance, both during readout and for the unique half-voltage RAM architecture, as well as retention ability. Reducing the thickness of the middle barrier causes energy splitting of the two QW ground states [274], which results in the QW_1 peak moving to a lower voltage requirement and the QW_2 peak moving to a higher voltage (Fig. 7.3**(c)**, red dots). There are no obvious disadvantages to this configuration, and the energy shift in QW_1 could be used for lower voltage memory operation using a single peak for the resonant-tunnelling current density. Reducing all barrier thicknesses results in a much broader current density relation, where it is likely that any gains in switching speed are outweighed by the degradation of other memory performance aspects. Most alterations possess desirable P cycle tunnelling characteristics for ULTRA**RAM**™ operation, which is positive for the purposes of production wafer tolerances. The performance trade-offs in the design have been identified, however, decisions on this should be informed after a more extensive experimental investigation of the target design.

The simulations are repeated with a reversed tunnelling bias (V_{TBRT}) to investigate the erase cycle of the memory (removing electrons from the FG). The current-density peaks which move electrons out of the FG occur at energy alignments with the QW_1 and QW_2 ground states in a similar fashion to the P cycle. This is demonstrated in Fig. 7.4 **(a)** and **(b)**, where the QW DOS alignments in the target structure correspond to the peaks as labelled in the current density plot (Fig. 7.4**(c)**, black squares). Increasing the barrier thicknesses reduces the magnitude of the current density peaks [Fig. 7.4**(c)**]. Increasing the thickness of all the barriers or the outer barriers (Fig. 7.4**(c)** green and blue triangles, respectively) reduces the contrast between peak current and off-current (orange shading), which could increase the cell logic disturbance rate. However, increased thickness of the middle barrier, B_2, (red dots) reduces peak current magnitude, but improves peak-off current ratio which would reduce logic disturbance of the memory in a RAM array [218]. Reducing barrier thicknesses increases the tunnelling current at the peaks, as shown in Fig. 7.4**(d)**. Reducing the outer barrier thickness makes little difference to the current-density relation, but shifts the peak current magnitude upwards by two orders (Fig 7.4**(d)**, green triangles). For the E cycle, the energy splitting from reducing the middle barrier thickness (Fig 7.4**(d)**, red dots) is less prominent, as biasing the structure in this direction moves the QW ground state energies further apart, thus reducing the interaction between them [274]. Moreover, reducing the thickness of the middle barrier (B_2) gains an order of magnitude in QW_1-peak current density and shifts it to a slightly lower voltage, whilst retaining similar off-currents as the target structure. Thus, the choice of reducing the middle barrier should be an improvement to the erase cycle tunnelling. However, these benefits should be carefully weighed against a potentially reduced retention capability and consequences of split P cycle tunnelling peaks, as previously discussed.

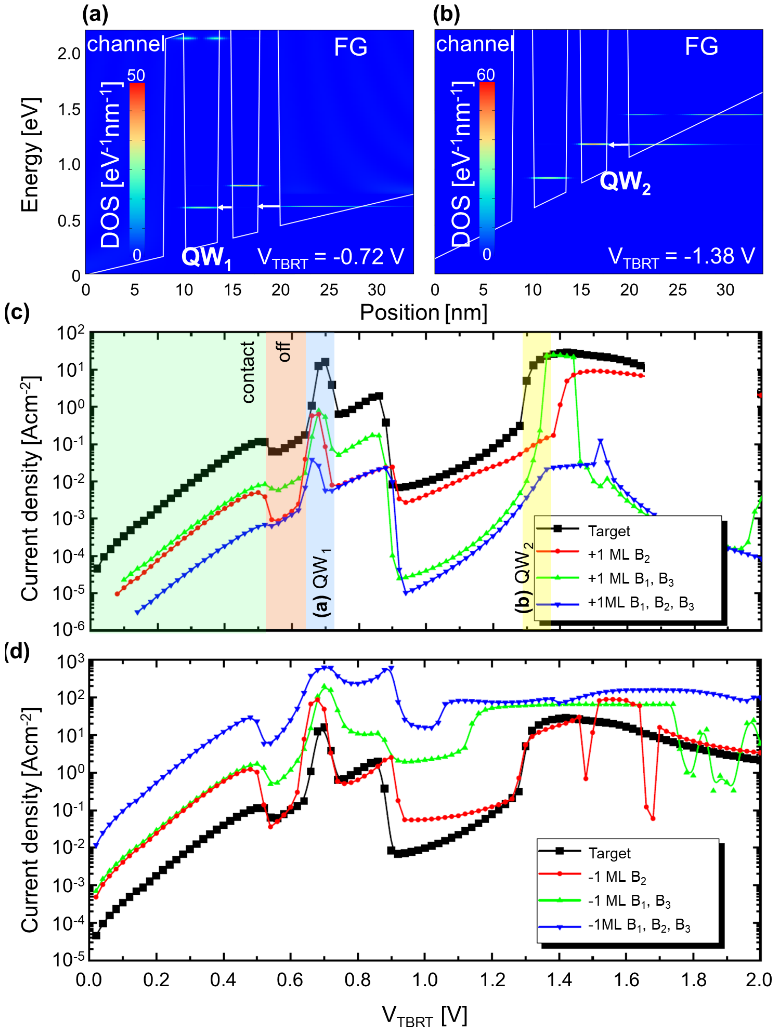


Figure 7.4: Nextnano MSB NEGF simulations (300 K) for the erase cycle. Results are described in detail in the text. **(a)** DOS for the QW_1 peak condition with FG-QW_1 energy alignment for resonant tunnelling for the target structure. **(b)** DOS for the QW_2 peak condition with FG-QW_2 energy alignment for resonant tunnelling for the target structure. **(c)** TBRT current density for increased barrier thicknesses compared with the target structure, where peak labels correspond to the alignments in **(a)** and **(b)**. **(d)** As in **(c)**, but for decreased barrier thicknesses. Figure reproduced from [273].

7.4 Summary

The InAs/AlSb TBRT region which forms the basis of the ULTRA**RAM**™ memory concept has been investigated in detail using the nextnano.MSB software package (NEGF with Büttiker probe scattering) to determine memory performance characteristics. Monolayer alterations are made to the tunnelling structure with the aim of realizing the optimum choice of layer structure whilst considering the growth tolerances required of a commercially-produced wafer. Transmission function calculations indicate that most monolayer alterations have a minimal effect on the retention capabilities. Thickening the barriers reduces low-energy transmission, improving retention, whilst the thinnest barrier configuration investigated allows 0.02 transmission at <300 meV, which could result in significant data (electron) losses. InAs QW widths can be engineered to alter the operation voltage for a specific purpose by shifting the ground state energies with little effect on the overall current density.

Trade-offs in current-density and retention are realized by comparison of zero-bias transmission with P and E cycle current density simulations. A higher current density will allow for high-speed operation, but retention may suffer as a result. As the upper limit of ULTRA**RAM**™'s speed and retention capabilities are still unknown (although early works show promising results), it is not possible to conclude which trade-off is more desirable for a non-volatile RAM. However, a promising candidate from this work is the reduction in the middle barrier thickness, where the current density is increased with little change to the zero-bias transmission function and program shifted to a lower voltage due to the QW_1-QW_2 ground state energy-splitting interaction. Moreover, the ability to tune the necessary P/E voltage by incremental changes to the tunnelling structure would allow for the technology to span many applications. Indeed, the optimal design for a DRAM replacement technology may prefer low-voltages and high-speed at the expense of retention, whereas an IoT sensor may prefer more robust retention at the expense of a slightly slower operation, for example. Crucially, monolayer changes to the TBRT region retain the unique physical phenomena which ULTRA**RAM**™ exploits for its superior performance metrics. As such, transfer of the technology from 3" Si [276] to commercial 12" Si is a real possibility if growth tolerances can be kept within ±1 ML (one lattice constant).

Chapter 8

Conclusions and Future Work

8.1 Conclusions

A novel FG-memory based on resonant tunnelling through InAs/AlSb heterostructures has been designed, modelled, fabricated and characterised in this thesis. Simulation works reveal the potential performance of ULTRA**RAM**™, predicting high-speed operation at a switching energy orders of magnitude lower than rival technologies. A unique channel design which allows high-contrast readout of memory cell logic is simulated. The design takes advantage of the unusual In(Ga)As/GaSb band offset such that a threshold-voltage can be obtained without the need for lateral doping [218]. A high-density architecture is proposed for RAM applications based on half-voltage P/E cycling, which is a design choice made possible by the unique tunnelling mechanism.

Memory devices and arrays are fabricated with their III-V layer designs based on the detailed understanding gained from prior simulations. The tunnelling direction is reversed compared to previous designs [216] and an ALD gate dielectric is introduced to significantly reduce leakage currents. The III-V memory layers were grown via MBE on GaAs and Si substrates where lattice-mismatch was alleviated using IMF layers. The fabrication process for single ULTRA**RAM**™ cells and 2×2 arrays was developed and optimised for different iterations of the technology on both substrate materials for 10, 20 and 50 µm gate lengths.

The outcome of both the theoretical and experimental investigations are summarised in Table 8.1. Crucially, ultra-low-power switching of non-volatile memory states is demonstrated for both GaAs and Si substrates. Table 8.1 compares ULTRA**RAM**™ performance metrics to other existing and emerging memory technologies, as previously presented in Section 2.4. However, these metrics are for devices at the 20 nm node size. In order to make a fair comparison with the experimental results of this thesis, metrics are converted to consider device area where it is relevant to the technology. An underlined metric in the table indicates that the performance has been demonstrated by experiment in this thesis, whereas non-underlined metrics indicates that the metric is from the simulations provided in Chapter 4. ULTRA**RAM**™ is a 1-transistor (1T) FG-memory technology. As there is only a single component required per bit, the potential for high-density memory arrays is promising, especially compared to other emerging memories which generally require two cell components per bit. However, aggressive scaling and significant improvements to logic readout contrast are required to produce the memory capacity to rival DRAM, for example.

ULTRA**RAM**™'s switching energy is incredibly low. At similar feature sizes, its energy requirement is lower than DRAM and flash by factors of 100 and 1000 respectively. As a FG-memory, the switching energy, E, is given by

$$E = \frac{1}{2}CV^2\ . \tag{8.1}$$

Consequently, capacitative scaling can be used to calculate the areal switching energy from the experimental results, which is again orders of magnitude lower than all current and emerging memory technologies. Non-volatility of the memory is confirmed over long periods ($>$24 hrs) at room temperature, and extrapolation of the data suggests a highly robust, non-volatile memory logic ($>$1000 yrs). Indeed, based on the barrier energy alone the logic retention time exceeds the age of the universe [284].

Memory endurance is confirmed experimentally for over 10^7 P/E cycles without any evidence of degradation. This represents a 10^5 improvement compared to the first prototype cells, which is most likely due to the elimination of gate leakage currents by inclusion of the ALD gate dielectric. Moreover, the endurance is at least two orders of magnitude improvement over flash memory, indicating that ULTRA**RAM**™ does not suffer from the degradation mechanisms that plague conventional FG memories. The limit of ULTRA**RAM**™'s endurance remains unknown due to experimental time constraints. Generally, there is a correlation between switching energy and memory endurance, suggesting that ULTRA**RAM**™ potentially has superior endurance properties owing to its low-energy P/E cycling. Small scale devices operating at higher speeds will offer an improved cycle frequency to investigate the upper endurance limit in a reasonable time frame.

The switching speed of the technology from simulation results is 500 ps, assuming that the input voltage on the device terminal has no delay from surrounding capacitances. Indeed, this metric should be thought of as a 'speed limit' based on the transport properties of the tunnelling structure. Experimentally, 500 µs switching is observed at a 20 µm feature size, which is due to the RC time constant from the large feature size. Assuming Dennard scaling holds to the 20 nm node, the areal speed is an order of magnitude lower than DRAM. This is expected from the intrinsically fast operation identified by simulation and the lower capacitance requirement of a FG memory compared to a DRAM cell.

The only apparent concern from the performance metrics of Table 8.1 is the particle number[1]. This is particularly noteworthy when scaling for area. Experimental threshold voltage shift (ΔV_T) measurements suggest a particle number that is a around six orders lower of magnitude

[1] Number of particles, in this case electrons, that define the memory logic.

than flash, corresponding to <19 electrons at the 20 nm node. This is less than one fifth of that predicted by simulations. However, this issue is related to the DOS in the FG layer and can be rectified by relatively straightforward layer alterations in later iterations, which would bring the value towards the prediction of the model.

Lastly, the TBRT structure, which is the source of ULTRA**RAM**™'s outstanding memory performance, is modelled in detail. The results indicate that monolayer alterations to the tunnelling layers have an impact on the device performance, and trade-offs relating to retention capability, switching speed and switching voltage are identified. Importantly, most monolayer changes to the TBRT region retain the physical properties required for ULTRA**RAM**™ operation. Consequently, feasibility of large scale manufacture is identified such that commercial 12" Si is a real possibility if growth tolerances can be reliably kept within one lattice constant.

Table 8.1: Benchmarking metrics for memory technologies [1, 36, 75, 92, 117] for ULTRA**RAM**™ comparison. Metrics that have been experimentally verified on large feature size devices (10-20 µm) are underlined. Metrics are otherwise simulated or scaled for 20 nm feature size.

Metric	**SRAM**	**DRAM**	**3D NAND-flash**	**PCM**	**ReRAM**	**FeRAM**	**STT-RAM**	ULTRA**RAM**™
Cell Area/F^2	120	6	<4	4-30	4-12	15-35	6	$<$**4**
Cell elements	6T	1T1C	1T	1T1R	1T1R	1T1C	1T1MTJ	**1T**
Voltage/V	< 1	< 1	>10	< 3	< 3	< 3	< 1.5	**2.5**
Switching energy/J	10^{-16}	10^{-15}	10^{-14}	10^{-10}	10^{-11}	10^{-11}	10^{-13}	$\mathbf{10^{-17}}$
Areal switching energy/J cm^{-2}	10^{-5}	10^{-4}	10^{-3}	-	-	-	-	$\mathbf{10^{-6}}$
Barrier energy/eV	-	0.5	1.6	2.4	1.4	-	1.5	**2.1**
Retention time	40 ms	60 ms	$>$ 10y	$>$ 10y	$>$ 10y	$>$ 10y	$>$ 10y	$>$ **10y***
Endurance	10^{16}	10^{16}	10^{5}	$10^{6} - 10^{13}$	$10^{6} - 10^{12}$	10^{12}	10^{12}	$\mathbf{\geq 10^{7}}$
Switching time	1 ns	10 ns	$>$ 10 µs	100-400 ns	10-100 ns	50 ns	10-50 ns	**500 ps/ 500 µs****
Areal switching time /s cm^{-2}	10^{2}	10^{3}	10^{6}	-	-	-	-	$\mathbf{10^{2}}$
Particle no.	-	10^{4}	$10^{3} - 10^{4}$	2×10^{4}	10-1000	-	10^{6}	**100**
Areal particle no./cm^{-2}	-	10^{19}	$10^{18} - 10^{19}$	10^{19}	$10^{16} - 10^{18}$	-	10^{21}	$\mathbf{10^{12}}$***

* Based on extrapolated experimental data. Temperature-accelerated testing is required to confirm this.

** 500 ps is the 1-dimension speed limitation from the simulation results and 500 µs is the switching speed measured from large devices.

*** Limited by electron DOS in the FG and can be increased with layer modifications.

8.2 Future work

In this thesis, a memory technology with unprecedented performance metrics has been presented and many of its attributes have been demonstrated experimentally on single devices and small arrays. However, there are substantial obstacles to overcome on the path to commercialisation. To function as a computer memory, many devices must be connected such that a series of binary digits can be stored to encode digital information. Currently, ULTRA**RAM**™ is limited to single-bit storage due to limitations in the readout contrast which has been identified as surface leakage from poor mesa isolation. Rectifying this issue will result in high-contrast memory readout which, in turn, allows for the development of large memory arrays in the proposed architecture. Fabrication and testing of large high-contrast memory arrays can be used for investigation of device uniformity and to assess feasibility of large scale integration.

Simultaneously, scaling the devices to the smallest node possible is a priority. The advantages of this are three-fold. Firstly, demonstrating the technology at small feature size gives some indication of potential bit-density once combined with the high-density architecture. Secondly, the upgrades to device performance with areal scaling can be demonstrated experimentally, verifying the claims of ultra-high-speed switching in this work. Thirdly, increasing device speed by scaling will allow for the P/E cycling frequency in endurance testing to be increased by orders of magnitude. Consequently, fabricating smaller devices is the most straightforward way of extending the endurance testing to approach RAM-suitable levels. Indeed, scaling to just 100 nm will extend the endurance cycling to beyond 10^{13} in within 100 hours.

Extended retention testing can be carried out using a temperature-accelerated method. This is the industry-standard way of testing the retention capability of FG memories ($>$ 10 yrs at room temperature) within a few weeks. However, the interpretation for temperature-accelerated data for an ULTRA**RAM**™ device will not be as straightforward due to the complexity of the tunnelling structure. The data analysis would require careful consideration of quantum effects when altering the distribution of carriers in the semiconductor layers by temperature increase, but may provide useful information about the long-term retention capabilities of the technology.

A problem arises when large, high-performance memory arrays of ULTRA**RAM**™ are successfully fabricated. Given the unusual materials used, designing the peripheral circuitry required to address arrays with many bitlines and wordlines is not straightforward. Conventionally, this is achieved with multiplexing using silicon CMOS logic circuitry, which, although possible, would be difficult to integrate into the III-V memory arrays. A more elegant solution is to create CMOS logic circuits from the III-V layers already present on the wafer for the memory. Naturally, the optimal ULTRA**RAM**™ channel design is of n-MOS construction, which could be appropriated for logic-use on the chip. Then, p-MOS operation could be obtained using the p-GaSb layers in the structure to complete a CMOS logic design for cell-address that fits seamlessly with the III-V memory design.

Lastly, the 3" Si wafer growth of the ULTRA**RAM**™ layer structure must be transferred to 12"

Si wafers whilst retaining layer quality and uniformity. This step will require significant efforts to implement, where metal-organic-chemical-vapour-deposition (MOCVD) can be used for growing the III-V layers instead of MBE. This reduces the cost of ULTRA**RAM**™ significantly, which is most important step towards convincing major semiconductor manufacturers that ULTRA**RAM**™ not only has superior performance, but is also a cost-effective, mass-manufacturable product.

Bibliography

[1] H.S. Wong and S. Salahuddin, ‘Memory leads the way to better computing,’ *Nat. Nanotechnol.*, vol. 10, 2015, pp. 191–194, 2015.

[2] J. Von Neumann, ‘First draft of a report on the EDVAC,’ *IEEE Annals of the History of Computing*, 15(4), pp. 27-75, 1993.

[3] H. Goldstine and American Council of Learned Societies, *The computer from Pascal to von Neumann (ACLS Humanities E-Book),* Princeton, N.J.: Princeton University Press, 1993.

[4] W.N. Toy and B. Zee, *Computer Hardware/Software Architecture,* Prentice Hall, 1986.

[5] J. Kang, *A Study of the DRAM Industry.* Master’s Thesis, Massachusetts Institute of Technology, Cambridge, MA, USA, 2010.

[6] G. E. Moore, ‘Progress in digital integrated electronics [Technical literature, Copyright 1975 IEEE. Reprinted, with permission. Technical Digest. International Electron Devices Meeting, IEEE, 1975, pp. 11-13.],’ *IEEE Solid-State Circuits Society Newsletter*, vol. 11, no. 3, pp. 36-37, 2006.

[7] R. Dennard, *et al*, ‘Design of ion-implanted MOSFETs with very small physical dimensions,’ *IEEE Journal of Solid State Circuits*, vol. SC-9, no.5, pp. 256-268, 1974.

[8] C. Carvalho, ‘The Gap between Processor and Memory Speeds,’ *ICCA,* 2002.

[9] C. H. Lam, ‘The Quest for the Universal Semiconductor Memory,’ *2005 IEEE Conference on Electron Devices and Solid-State Circuits,* Howloon, Hong Kong, pp. 327-331, 2005.

[10] J. Akerman, ‘Toward a Universal Memory,’ *Science,* Vol. 308, Issue 5721, pp. 508-510, 2005.

[11] N. Jones, ‘How to stop data centres from gobbling up the world’s electricity,’ *Nature 561,* pp. 163-166, 2018.

[12] Greenpeace, *Clicking clean: who is winning the race to build a green internet?,* 2017. [available online: clickclean.org/international/en]

[13] J. Truby, ‘Decarbonizing Bitcoin: Law and Policy Choices for Reducing the Energy Consumption of Blockchain Technologies and Digital Currencies,’ *Energy Research & Social Science,* vol. 44, pp. 399–410, 2018.

[14] L. Barroso, J. Clidaras, and U. Holzle, *The datacenter as a computer an introduction to the design of warehouse-scale machines (2nd ed., Synthesis digital library of engineering and computer science),* San Rafael, Calif.: Morgan & Claypool, 2013.

[15] A. Shehabi, B. Walker, and E. Masanet, ‘The energy and greenhouse-gas implications of internet video streaming in the United States,’ *Environmental Research Letters,* 9(5), pp. 1-11, 2014.

[16] E. Brynjolfsson, and A. McAfee, *The second machine age : Work, progress, and prosperity in a time of brilliant technologies,* New York: W.W. Norton & Company, 2014.

[17] K. Schwab, *The Fourth industrial revolution,* United Kingdom: Portfolio Penguin, 2017.

[18] A. Sebastian, ‘Computational memory: A stepping-stone to non-von Neumann computing?,’ *Stanford EE Computer Systems Colloquium,* 2018. [available online: http://web.stanford.edu/class/ee380/Abstracts/180307.html]

[19] L. Null, J. Lobur *et al*, *The essentials of computer organization and architecture [fourth edition],* Jones & Bartlett Publishers, 2014.

[20] M.L. Page, ‘AI’s dirty secret: Energy-guzzling machines may fuel global warming,’ *Newscientist,* Magazine issue 3199 , published 13 October 2018.

[21] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, ‘Internet of Things (IoT): A vision, architectural elements, and future directions,’ *Future generation computer systems,* 29(7), pp. 1645-1660, 2013.

[22] Memory. In Oxford Dictionary. [Available online: https://www.lexico.com/en/definition/memory]

[23] Jean-Victor Poncelet, ‘Travaux de la Commission Francaise,’ *L’Exposition Universelle de 1851,* vol. 3, part 1 (Machines et outils appliques aux arts textiles), section 2, pages 348-349, 1857.

[24] C. Shannon, ‘A symbolic analysis of relay and switching circuits,’ *Electrical Engineering,* 57(12), 713-723, 1938.

[25] C. Shannon, ‘A Mathematical Theory of Communication,’ *Bell System Technical Journal,* 27(3), 379-423, 1948.

[26] D. Laws, ‘A Company of Legend: The Legacy of Fairchild Semiconductor,’ *IEEE Annals of the History of Computing,* vol. 32, no. 01, pp. 60-74, 2010. doi: 10.1109/MAHC.2010.12

[27] Intel Corporation, ‘A chronologic list of Intel products. The products are sorted by date,’ Intel museum, 2005.

[28] A.G.F. Dingwall and R. E. Stieker, ‘Compact COS/MOS 256-bit random access memory,’ *1970 International electron devices meeting (IEDM),* pp. 101-103, 1974.

[29] Intel Corporation, ‘Silicon Gate MOS 1101A, 1101A1, 256 bit fully decoded random access memory (datasheet)’ *Memory Intel Vintage,* 1101 [available at: intel-vintage.info/intelmemory.org].

[30] O. Kenneth, H. Best and L. Richard. *Magnetic Core Memory,* 1964.

[31] S.M. Sze, *Semiconductor devices: Physics and technology,* Wiley 2nd ed., 2002.

[32] S.M. Sze, *Modern Semiconductor Device Physics,* New York: Wiley, 1998.

[33] S. Jones, ‘7nm, 5nm and 3nm Logic, current and projected processes,’ *SemiWiki,* 2019.

[34] D. Schor, ‘TSMC Starts 5-Nanometer Risk Production,’ *WikiChip Fuse* [Retrieved 2019-04-07].

[35] AKM, ‘Japanese Company Profiles’ *Smithsonian Institution,* 1993. [available at: http://smithsonianchips.si.edu/ice/cd/PROF96/JAPAN.PDF]

[36] S. Yu and P.Y. Chen ‘Emerging memory technologies: recent trends and prospects’ *IEEE Solid-State Circuits Magazine,* 8(2):43-56, 2016.

[37] S. Adee, ‘Thanks for the Memories,’ *IEEE Spectrum,* 2009.

[38] R. H. Dennard, ‘Field effect transistor memory,’ US Patent US3387286A, 1968.

[39] R. H. Dennard, ‘Revisiting Evolution of the MOSFET Dynamic RAM – A Personal View,’ *IEEE Solid-State Circuits Society Newsletter,* vol. 13, no. 1, pp. 10-16, 2008.

[40] ‘JEDEC Double Data Rate (DDR) SDRAM Specification,’ JESD79C. *JEDEC Solid State Technology Assoc,* March 2003. p.20, on School of Engineering and Computer Science, Baylor Univ. website, 2003.

[41] V.M. Shikhare and S.Oza, ‘Reducing Power Consumption in DRAM Using Partial Access Method,’ *IJEEDC,* ISSN (P): 2320-2084, (O) 2321–2950, 2015.

[42] Y. Li, *Robust Design of DRAM Core Circuits - Yield Estimation and Analysis by A Statistical Design Approach,* PhD thesis, TECHNISCHE UNIVERSITÄT MÜNCHEN 2010.

[43] Lecture 12: DRAM Basics (PDF). utah.edu. 2011-02-17. Archived (PDF) [available at: https://web.archive.org/web/20150616050009/http://www.eng.utah.edu/ cs7810/pres/11-7810-12.pdf. Accessed 26/4/21.]

[44] S. Narasimha *et al*, ‘22nm High-performance SOI technology featuring dual-embedded stressors, Epi-Plate High-K deep-trench embedded DRAM and self-aligned Via 15LM BEOL,’ *2012 International Electron Devices Meeting,* San Francisco, CA, pp. 3.3.1-3.3.4, 2012. doi: 10.1109/IEDM.2012.6478971

[45] P. Horowitz and W. Hill, *The art of electronics (2nd ed.),* Cambridge: Cambridge University Press, 1989.

[46] H. Sunami, ‘The Role of the Trench Capacitor in DRAM Innovation,’ *IEEE Solid-State Circuits Society Newsletter,* vol. 13, no. 1, pp. 42-44, 2008. doi: 10.1109/N-SSC.2008.4785691

[47] M. Gutsche, H. Seidl, T. Hecht, S. Kudelka and U. Schroeder, ‘Atomic Layer Deposition for advanced DRAM applications,’ *Future Fab International,* 15, 2003.

[48] R. W. Johnson, A.Hultqvist and S. F. Bent, ‘A brief review of atomic layer deposition: from fundamentals to applications,’ *Materials Today,* vol. 17, Issue 5, p. 236-246, 2014.

[49] G. Fredeman *et al*, ‘17.4 A 14nm 1.1Mb embedded DRAM macro with 1ns access,’ *2015 IEEE International Solid-State Circuits Conference* - (ISSCC) Digest of Technical Papers, San Francisco, CA, pp. 1-3, 2015. doi: 10.1109/ISSCC.2015.7063053

[50] R. Bez, E. Camerlenghi, A. Modelli and A. Visconti, ‘Introduction to Flash memory,’ *Proceedings of the IEEE,* 91. 489 - 502, 2003. 10.1109/JPROC.2003.811702

[51] S. S. Haddad and Hao Fang. ‘Method for bulk (or byte) charging and discharging an array of flash EEPROM memory cells,’ *U.S. Patent No. 5,491,657* , 1996.

[52] D. Richter, *Flash Memories: Economic Principles of Performance, Cost and Reliability,* Springer Science and Business Media. pp. 5–6, 2013. doi:10.1007/978-94-007-6082-0. ISBN 978-94-007-6081-3.

[53] T. Shibata and T. Ohmi, ‘A functional MOS transistor featuring gate-level weighted sum and threshold operations,’ *IEEE Transactions on Electron Devices,* vol. 39, no. 6, pp. 1444–1455, 1992.

[54] B. Kalyan, and B. Singh, ‘Design and simulation equivalent model of Floating Gate Transistor,’ India Conference (INDICON), 2015 Annual IEEE, pp. 1-6, 2015.

[55] R. Bez and A. Pirovano, *Advances in Non-Volatile Memory and Storage Technology,* Woodhead Publishing, 2019. ISBN 9780081025857.

[56] C. Zhao, C. Z. Zhao and S. Taylor, P. R. Chalker, ‘Review on Non-Volatile Memory with High-k Dielectrics: Flash for Generation Beyond 32 nm,’ *Materials,* 7, 5117-5145, 2014. doi:10.3390/ma7075117

[57] S. Maikap *et al*, 'Enhanced flash memory device characteristics using ALD TiN/Al2O3 nanolaminate charge storage layers,' *2008 9th International Conference on Solid-State and Integrated-Circuit Technology,* Beijing, pp. 958-961, 2008. doi: 10.1109/IC-SICT.2008.4734702

[58] G.Chen, Z. Huo, S. Zhao, X. Yang, Z. Liu, M. Zhang, Z. Sun, Y. Han, D. Zhang, C. Wang and Y. Chu, 'Optimization of HfO2 Growth Process by Atomic Layer Deposition (ALD) for High Performance Charge Trapping Flash Memory Application,' *ECS Transactions,* 2013 Mar 8;52(1):51-6, 2013.

[59] TSMC *eFlash* Taiwan Semiconductor Manufacturing Company (TSMC) webpage on eFlash. [Available at https://www.tsmc.com/english/dedicatedFoundry/technology/eflash.htm eFlash. Accessed 26/04/2021.]

[60] R. H. Fowler and L. Nordheim, 'Electron emission in intense electric fields,' *119 Proc. R. Soc. Lond. A,* 1928. http://doi.org/10.1098/rspa.1928.0091

[61] M. Lenzlinger and E.H Snow. 'Fowler-Nordheim Tunneling into Thermally Grown SiO2,' *IEEE Transactions on Electron Devices,* vol. 15, no. 9, pp. 686, 1968.

[62] A. Kolodny, S. T. K. Nieh, B. Eitan and J. Shappir, 'Analysis and modeling of floating-gate EEPROM cells,' *IEEE Transactions on Electron Devices,* vol. 33, no. 6, pp. 835-844, 1986. doi: 10.1109/T-ED.1986.22576

[63] S. S. Chung, C.-M. Yih, S. S. Wu, H. H. Chen, and G. Hong, 'A SPICE-compatible Flash EEPROM model feasible for transient and program/erase cycling endurance simulation,' *in IEDM Tech. Dig.,* pp. 179–182, 1999.

[64] A. Grinberg, A. Kastalsky and S. Luryi, 'Theory of hot-electron injection in CHINT/NERFET devices,' *IEEE Transactions on Electron Devices,* 34(2), 409-419, 1987.

[65] A. Fazio. 'Flash memory scaling,' *MRS Bull,* 29:814–817, 2004.

[66] F. Hsu and H. R. Grinolds, 'Structure-enhanced MOSFET degradation due to hot-electron injection,' *in IEEE Electron Device Letters,* vol. 5, no. 3, pp. 71-74, 1984. doi: 10.1109/EDL.1984.25836

[67] A. S. Gorobets *et al*, 'Cyclic flash memory wear leveling,' *U.S. Patent No. 7,441,067*, 2008.

[68] Seong-kue Jo, 'Flash memory device for performing bad block management and method of performing bad block management of flash memory device,' *U.S. Patent No. 7,434,122*, 2008.

[69] Samsung SSD 840 evo specification. [Available at: http://www.samsung. com/global/business/semiconductor/ minisite/SSD/global/html/about/ SSD840EVO.html.]

[70] J. Meza *et al*, 'A large-scale study of flash memory failures in the field,' *ACM SIGMETRICS Performance Evaluation Review,* Vol. 43. No. 1. ACM, 2015.

[71] S. S. Chung *et al*, 'A new technique for hot carrier reliability evaluations of flash memory cell after long-term program/erase cycles,' *IEEE Transactions on Electron Devices,* 46.9: 1883-1889, 1999.

[72] J. J. Chen, N. R. Mielke and C. C. Hu, 'Nonvolatile Memory Technologies with Emphasis on Flash,' *IEEE Press Series on Microelectronic Systems,* John Wiley & Sons, Inc., Hoboken, NJ, USA, 2007.

[73] H. W. You and W. J. Cho, W. J, 'Charge trapping properties of the HfO 2 layer with various thicknesses for charge trap flash memory applications,' *Applied Physics Letters*, 96(9), 093506, 2010.

[74] U. Satya Sainadh, Han Xu, Xiaoshan Wang, A. Atia-Tul-Noor, William C. Wallace, Nicolas Douguet, Alexander Bray, Igor Ivanov, Klaus Bartschat, Anatoli Kheifets, R. T. Sang and I. V. Litvinyuk, 'Attosecond angular streaking and tunnelling time in atomic hydrogen,' *Nature,* volume 568, p.75–77, 2019.

[75] K. Prall, 'Benchmarking and Metrics for Emerging Memory,' 2017 IEEE International Memory Workshop (IMW), Monterey, pp. 1-5, 2017.

[76] O. Wong, Hei Wong, Wing-Shan Tam and Chi-Wah Kok, 'An overview of charge pumping circuits for flash memory applications,' *2011 9th IEEE International Conference on ASIC,* Xiamen, pp. 116-119, 2011. doi: 10.1109/ASICON.2011.6157136

[77] J. A. Starzyk, Ying-Wei Jan, and Fengjing Qiu. 'A DC-DC charge pump design based on voltage doublers,' *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications* 48.3, 350-359, 2001.

[78] Micron, 'TN-29-07: Small-Block vs. Large-Block NAND Flash Devices', 2005. [Available at: https://www.micron.com/support/ /media/74C3F8B1250D4935898DB7FE79EB56E7.ashx].

[79] V. Mohan, *Modeling the physical characteristics of NAND flash memory,* Ph.D. dissertation, Univ. Virginia, Charlottesville, VA, USA, 2010.

[80] Tae-Sung Jung, Young-Joon Choi, Kang-Deog Suh, Byung-Hoon Suh, Jin-Ki Kim, Young-Ho Lim, Yong-Nam Koh, Jong-Wook Park, Ki-Jong Lee, Jung-Hoon Park, Kee-Tae Park, Jhang-Rae Kim, Jeong-Hyong Yi and Hyung-Kya Lim, 'A 117-mm2 3.3-V only 128-Mb multilevel NAND flash memory for mass storage applications. Solid-State Circuits,' *IEEE Journal of,* 31. 1575 - 1583, 1996. 10.1109/JSSC.1996.542301.

[81] Jagan Singh Meena, Simon Min Sze, Umesh Chand and Tseung-Yuen Tseng 'Overview of emerging nonvolatile memory technologies,' *Nanoscale Research Letters* volume 9, Article number: 526, 2014.

[82] R. Klanderman, *Flash Memory Device: Electrical Modeling and Simulation,* MSc thesis, TU Delft, Netherlands, 2012.

[83] Micron, *NOR|NAND Flash Guide* [Available at: https://www.micron.com/-/media/client/global/documents/products/product-flyer/nor_nand_flash_guide.pdf?la=en].

[84] J. Maimon, E. Spall, R. Quinn and S. Schnur, 'Chalcogenide-based non-volatile memory technology,' *2001 IEEE Aerospace Conference Proceedings* (Cat. No.01TH8542), pp. 2289-2294 vol.5, 2001. doi: 10.1109/AERO.2001.931188.

[85] H.S.P. Wong, S. Raoux, S. Kim, J. Liang, J.P. Reifenberg, B. Rajendran and K.E. Goodson, 'Phase change memory,' *Proceedings of the IEEE*, 98(12), 2201-2227, 2010.

[86] L. Jiang, B. Zhao, Y. Zhang, J. Yang and B.R. Childers, 'Improving write operations in MLC phase change memory,' *IEEE International Symposium on High-Performance Comp Architecture*, pp. 1-10, 2012.

[87] enquoteIntel Launches Optane Memory M.2 Cache SSDs for Consumer Market, AnandTech, 2017. [Available online: Retrieved 13 November 2017 https://www.anandtech.com/show/11227/intel-launches-optane-memory-m2-cache-ssds-for-client-market].

[88] 'Hey, Intel and Micron: XPoint is phase-change memory, right? Or is it? Yes. No. Yes', 2016. [Available online: https://www.theregister.co.uk/2016/01/19/xpoint_intel_micron_phasechange].

[89] Z. Song, S. Song, M. Zhu, L. Wu, K. Ren, W. Song and S. Feng, 'From octahedral structure motif to sub-nanosecond phase transitions in phase change materials for data storage,' *Science China Information Sciences*, 61(8), 081302, 2018.

[90] [Available online: http://ovonyx.com/technology/technical-presentation.html.]

[91] G. Servalli, 'A 45nm generation phase change memory technology,' *IEEE IEDM*, 2009.

[92] W. Banerjee, 'Challenges and Applications of Emerging Nonvolatile Memory Devices,' *Electronics* 9, 1029, 2020.

[93] A. Pirovano, 'An Introduction on Phase-Change Memories,' *Redaelli A. (eds) Phase Change Memory.* Springer, Cham, 2018.

[94] T. Ninimiya, Z. Wei, S. Muraoka, R. Yasuhara, K. Katayama and T. Takagi. 'Conductive filament scaling of TaOx bipolar ReRAM for improving data retention under low operation current,' *IEEE Transactions on Electron Devices,* 60(4): 1384-1389, 2013.

[95] Y. Hayakawa, 'High reliable TaOx ReRAM with centralized filament for 28nm embedded application,' *2015 IEEE VLSI,*, 2015.

[96] S. Ambrogio *et al*, ‘Statistical fluctuations in HfOx resistive-switching memory: part II - random telegraph noise,’ *IEEE TRED*, vol. 61, no. 8, pp. 2920-2927, 2014.

[97] A. Levisse, P. Royer, Bastien Giraud, J.P. Nöel, Mathieu Moreau, *et al,* ‘Architecture, design and technology guidelines for crosspoint memories,’ *2017 IEEE/ACM International Symposium on Nanoscale Architectures (NANOARCH),* Newport, United States. pp.677-686, 2017.

[98] M. Soni and R. Dahiya, ‘Soft eSkin: distributed touch sensing with harmonized energy and computing,’ *Philosophical Transactions of the Royal Society of London A,* 378(2164), 2020.

[99] K-C. Kwon, M-J. Song, K. Kwon, H.V. Jeong, D.W. Kim, G-S. Lee, J-P. Hong and J-G. Park, ‘Nanoscale CuO solid-electrolyte-vased conductive-bridging-random-access memory cell operating multi-level-cell and 1selector1resistor,’ *Journal of Materials Chemistry C* 3(37):9540-9550, 2015.

[100] M. Kund, G. Beitel, C.U. Pinnow, T. Rohr, J. Schumann, R. Symanczyk and G. Muller, ‘Conductive bridging RAM (CBRAM): An emerging non-volatile memory technology scalable to sub 20nm,’ *IEEE International Electron Devices Meeting,* IEDM Technical Digest, pp. 754-757, 2005.

[101] A. Calderoni *et al*, ‘Performance comparison of O-based and Cu-based ReRAM for high-density applications,’ *IEEE IMW 2014*, 2014.

[102] N. Gonzales *et al*, ‘An Ultra low-power non-volatile memory design subquantum conductive-bridge RAM,’ *2016 IEEE IMW*, 2016.

[103] D.Jana, S. Roy, R. Panja, M. Dutta, S.Z. Rahaman, R. Mahapatra and S. Maikap, ‘Conductive-bridging random access memory: challenges and opportunity for 3D architecture,’ *Nanoscale research letters*, 10, 188, 2015. doi:10.1186/s11671-015-0880-9

[104] History of FRAM, [available online: https://www.fujitsu.com/downloads/MICRO/fme/fram/fram-guide-book.pdf].

[105] P.K. Larsen, R. Cuppens and G. Spierings. ‘Ferroelectric memories,’ *Ferroelectrics* 128(1): 265-292, 1992.

[106] Mjitsu Microelectronics. *FRAM guide book,* Fjitsu, 2008.

[107] N. Inoue, Y. Maejima and Y. Hayashi, ‘Crystal-orientation controlled PZT FeRAM-capacitors using RF magnetron sputtering with 12"/spl phi/ single target,’ *International Electron Devices Meeting,* IEDM Technical Digest, Washington, DC, USA, pp. 605-608, 1997. doi: 10.1109/IEDM.1997.650457

[108] T. Endoh, H. Koike, S. Ikeda, T. Hanyu and H. Ohno, ‘An overview of nonvolatile emerging memories—Spintronics for working memories,’ *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 6(2), 109-119, 2016.

[109] L. Grenouillet *et al.*, ‘Performance assessment of BEOL-integrated HfO2-based ferroelectric capacitors for FeRAM memory arrays,’ *2020 IEEE Silicon Nanoelectronics Workshop (SNW)*, pp. 5-6, 2020. doi: 10.1109/SNW50361.2020.9131648.

[110] A. Chen, ‘A review of emerging non-volatile memory (NVM) technologies and applications,’ *Solid-State Electronics,* vol.125, pp. 25-38, 2016.

[111] H. Mulaosmanovic, E. T. Breyer, T. Mikolajick and S. Slesazeck, ‘Ferroelectric FETs With 20-nm-Thick HfO2 Layer for Large Memory Window and High Performance,’ *IEEE Transactions on Electron Devices,* vol. 66, no. 9, pp. 3828-3833, 2019. doi: 10.1109/TED.2019.2930749.

[112] J.C. Slonczewski, ‘Current-driven excitation of magnetic multilayers,’ *Journal of Magnetism and Magnetic Materials*, vol. 159, Issues 1–2, Pages L1-L7, 1996. https://doi.org/10.1016/0304-8853(96)00062-5.

[113] H. Yoda, ‘MRAM Fundamentals and Devices’ *Handbook of Spintronics,* Springer, Dordrecht, 2015.

[114] D. Denny and Yuan-Jen Lee Tang, *Magnetic memory : fundamentals and technology,* New York : Cambridge University Press, 2010.

[115] S. Li, ‘Spin Transfer Torque-RAM Devices as a Future Non-volatile Memory Solution,’ [Available online: https://pdfs.semanticscholar.org/6c6f/68402ce28cd13cc2f5b96a3684fd82807b30.pdf]

[116] D. Tang *et al*, *Magnetic Memory - Fundamentals and Technology* Cambridge Univ. Press, pp. 122-165, 2010.

[117] C. Cakir, M. Bhargava and K. Mai, ‘6T SRAM and 3T DRAM data retention and remanence characterization in 65nm bulk CMOS,’ *Proceedings of the IEEE 2012 Custom Integrated Circuits Conference,* San Jose, CA, pp. 1-4, 2012. doi: 10.1109/CICC.2012.6330672

[118] Electron Microscopy Tutorial, University of Utah. [Available online: https://advanced-microscopy.utah.edu/education/electron-micro/]

[119] O. C. Wells, *Scanning Electron Microscopy,* McCraw-Hill, New York, 1974.

[120] G. Lawes *et al*, *Scanning Electron Microscopy and x-Ray Microanalysis,* Published on Behalf of ACOL, Thames Polytecnic, London, by Wiley, 1987.

[121] G. K. Benning, ‘Atomic force microscope and method for imaging surfaces with atomic resolution,’ US Patent US4724318A, 1986.

[122] *All Nobel Prizes in Physics,* Nobelprize.org. Nobel Media AB.

[123] E. Meyer, J. Hug and R. Bennewitz, *Scanning Probe Microscopy,* Advanced Texts in Physics, Springer Berlin Heidelberg, 2004.

[124] G. Haugstad, *Atomic Force Microscopy: Understanding Basic Modes and Advanced Applications,* John Wiley & Sons, 2012.

[125] A. Robson, I. Grishin, R. J. Young, A. M. Sanchez, O. V. Kolosov, and M. Hayne, 'High-Accuracy Analysis of Nanoscale Semiconductor Layers Using Beam-Exit Ar-Ion Polishing and Scanning Probe Microscopy,' *ACS Applied Materials & Interfaces,* 3241-3245, 2013.

[126] J. Fruhauf, 'Problems of contour measuring on microstructures using a surface profiler,' *Measurement Science and Technology,* 9(3): 293-296, 1998.

[127] Jiunn-Jong Wu, 'Spectral analysis for the effects of stylus tip curvature on measurement iotropic rough surfaces,' *Measurement Science and Technology,* 13(5):310, 2002.

[128] A. Holland, *GENERAL TRAINING MANUAL FOR JOBIN YVON HORIBA THIN FILM GROUP END POINT DETECTION EQUIPMENT,* Jobin Yvon Ltd., Horiba Group, 1999.

[129] T. Wilson. *The Design, Optimisation, and Characterisation of GaSb/GaAs Quantum Ring-Based Vertical-Cavity Devices Emitting at Telecoms Wavelengths,* PhD Thesis, Lancaster University 2021.

[130] J.W. Matthews. *Epitaxial Growth,* Academic Press, 1975.

[131] R. Farrow, *Molecular Beam Epitaxy,* New York: Elsevier Science & Technology, 2014.

[132] W. Braun, *Applied RHEED: Reflection High-Energy Electron Diffraction During Crystal Growth,* Springer-Verlag: Berlin, 1999.

[133] J.H. Neave, B.A. Joyce and P.J. Dobson, 'Dynamic RHEED observations of the MBE growth of GaAs,' *Applied Physics A Solids and Surfaces,* 34(3): 179-184, 1984.

[134] L.F. Thompson and R. E. Kerwin. 'Polymer resist systems for photo-and electron lithography,' *Annual Review of Materials Science* 6.1: 267-301, 1976.

[135] V. Bakshi. *EUV lithography,* Bellingham, Wash. : Hoboken, NJ: SPIE Press ; John Wiley, 2009.

[136] B. Thedjoisworo, D. Cheung, and V. Crist, 'Comparison of the effects of downstream H2- and O2-based plasmas on the removal of photoresist, silicon, and silicon nitride,' *J. Vac. Sci. Technol. B* 31, 021206, 2013. https://doi.org/10.1116/1.4792254,

[137] CK Chung, *Plasma Etching,* In: Li D. (eds) Encyclopedia of Microfluidics and Nanofluidics. Springer, Boston, MA, 2014.

[138] D.M. Mattox, 'The Low Pressure Plasma Processing Environment,' *Handbook of Physical Vapour Deposition (PVD) Processing,* pap. 157-193. Elsevier, 2010.

[139] G.S. May and C.J. Spanos, *Fundamentals of Semiconductor Manufacturing and Process Control,* IEEE ; Wiley-Interscience, 2006.

[140] T.S. Chao, *Introduction to semiconductor manufacturing technology,* SPIE PRESS, 2001.

[141] J.E. Mahan, 'Physical vapor deposition of thin films,' *Physical Vapor Deposition of Thin Films, by John E. Mahan,* pp. 336. ISBN 0-471-33001-9. Wiley-VCH, 2000.

[142] J. Handley, 'Product Review: Quartz Crystal Microbalances,' *ACS Publications*: 225-A, 2001.

[143] Z. Wang and Z. Zhang, 'Electron beam evaporation deposition,' *Advanced Nano Deposition Methods,* Wiley-VCH Verlag GmbH & Co. KGaA; pp. 33-58, 2016.

[144] Kurt J Lesker Company, Material Deposition Chart. [Available online: https://www.lesker.com/newweb/deposition_materials/materialdepositionchart.cfm?pgid=0]

[145] W.B. Hanson, S. Sanatani, and J. H. Hoffman, 'Ion sputtering from satellite surfaces,' *J. Geophys. Res.,* 86, 11350- 11356, 1981.

[146] P.D. Davidse and L. I. Maissel. 'Dielectric thin films through rf sputtering,' *Journal of Applied Physics* 37.2: 574-579, 1966.

[147] T. Weckman and K. Laasonen, 'First principles study of the atomic layer deposition of alumina by TMA–H2O-process,' *Phys. Chem. Chem. Phys.,* 17, 17322-17334, 2015.

[148] N. Batra, J. Gope, Vandana, J. Panigrahi, R. Singh and P. K. Singh, 'Influence of deposition temperature of thermal ALD deposited Al2O3 films on silicon surface passivation,' *AIP Advances 5,* 067113, 2015. https://doi.org/10.1063/1.4922267

[149] M.D. Halls and K. Raghavachari, 'Atomic Layer Deposition Growth Reactions of Al 2 O 3 on Si(100)-2×1,' *Journal of Physical Chemistry B,* 108, 4058– 4062, 2004. DOI: 10.1021/jp0378079

[150] M. Groner, F. Fabreguette, J. Elam and S. George, 'Low-temperature Al2O3 atomic layer deposition,' *Chemistry of Materials*, 16 (4), 639-645. (50), 2004.

[151] S. M. Prokes, M. B. Katz and M. E. Twigg 'Growth of crystalline Al2O3 via thermal atomic layer deposition: Nanomaterial phase stabilization,' *APL Materials 2,* 032105, 2014. https://doi.org/10.1063/1.4868300

[152] R. W. Johnson, A. Hultqvist, S. F. Bent, 'A brief review of atomic layer deposition: from fundamentals to applications,' *Materials Today,* Volume 17, 5 pp 235-246, 2014. https://doi.org/10.1016/j.mattod.2014.04.026.

[153] J.S. Lim, *et al*, *ICVC '99 6th International Conference,* pp. 506-509, 1999.

[154] K. Mistry *et al*, 'A 45nm Logic Technology with High-k+ Metal Gate Transistors, Strained Silicon, 9 Cu Interconnect Layers, 193nm Dry Patterning, and 100% Pb-free Packaging,' *2007 IEEE International Electron Devices Meeting,* Washington, DC, pp. 247-250, 2007. doi: 10.1109/IEDM.2007.4418914

[155] S. Hyun *et al*, 'Aggressively scaled high-k last metal gate stack with low variability for 20nm logic high performance and low power applications,' *2011 Symposium on VLSI Technology-Digest of Technical Papers,* pp. 32-33, 2011.

[156] S. Narasimha *et al*, '22nm High-performance SOI technology featuring dual-embedded stressors, Epi-Plate High-K deep-trench embedded DRAM and self-aligned Via 15LM BEOL,' *2012 International Electron Devices Meeting,* San Francisco, CA, pp. 3.3.1-3.3.4, 2012. doi: 10.1109/IEDM.2012.6478971

[157] F. Koehler, D. H. Triyoso, I. Hussain, S. Mutas and H. Bernhardt, 'Atomic Layer Deposition of SiN for spacer applications in high-end logic devices,' *IOP Conference Series: Materials Science and Engineering,* vol. 41(1), p. 012006, 2012.

[158] C. Auth *et al*, 'A 22nm high performance and low-power CMOS technology featuring fully-depleted tri-gate transistors, self-aligned contacts and high density MIM capacitors,' *2012 Symposium on VLSI Technology* (VLSIT) p.131-132, 2012.

[159] R.C.G. Swann, 'The Birth of Glow Discharge Chemistry' *Engineering and Technology History Wiki* [Available online: https://ethw.org/w/index.php?title=First-Hand:The_Birth_of_Glow_Discharge_Chemistry_(aka_PECVD)&oldid=116850. (2015). Accessed 26/04/21.]

[160] S.E. Alexandrov and M.L. Hitchman. 'Chapter 12: Plasma-enhanced Chemical Vapour Deposition Processes,' *Chemical Vapour Deposition,* pp 494-534, Royal Society of Chemistry, Cambridge, 2009.

[161] L. Marinu, O. Zbeida and J.E. Klemberg-Sapeha.'Plasma-Enhanced Chemical Vapour Deposition of Functional Coatings' *Handbook of Deposition Technologies for Films and Coatings,* pp. 392-465, Elsevier, 2010.

[162] Spectrum DIP datasheet, [Available online: https://www.spectrum-semi.com/CSB02813.pdf]

[163] S. Birner, nextnano++ software documentation. [Available online: https://www.nextnano.de/nextnanoplus/]

[164] S. Birner, *Modeling of Semiconductor Nanostructures and Semiconductor-electrolyte Interfaces,* PhD dissertation, Technischen Universitat Muchen, 2011.

[165] P. Greck, S. Birner, B. Huber, and P. Vogl, 'Efficient method for the calculation of dissipative quantum transport in quantum cascade lasers,' *Opt. Express 23,* 6587-6600, 2015.

[166] S. Birner and T. Grange, nextnano.MSB software documentation. [Available online: https://www.nextnano.de/nextnano3/nextnano.MSB/index.htm]

[167] P. Greck, 'Efficient calculation of dissipative quantum transport properties in semiconductor nanostructures,' *Selected Topics of Semiconductor Physics and Technology* (G. Abstreiter, M.-C. Amann, M. Stutzmann, and P. Vogl, eds.), vol. 105, Verein zur Förderung des Walter Schottky Instituts der Technischen Universität München e.V., München, 2012.

[168] R. Yatskiv and J. Voves, 'Analysis of the resonant tunneling diode with the stepped prebarrier,' *Journal of Phyics: Conference Series,* 193 012007, 2009.

[169] A. Vladimirescu, *The Spice book,* New York: J. Wiley, 1994.

[170] LTspice software documentation. [Available online: https://www.analog.com/en/design-center/design-tools-and-calculators/ltspice-simulator.html].

[171] IUPAC Periodic Table of the Elements (2016). [Available online at: https://iupac.org/wp-content/uploads/2015/07/IUPAC_Periodic_Table-28Nov16.pdf]

[172] Zeke Liu, Wanli Ma, Xingchen Ye, 'Chapter 2 : Shape control,' *the synthesis of colloidal semiconductor nanocrystals in Anisotropic Particle Assemblies,* Elsevier, pp. 37-54, 2018. https://doi.org/10.1016/B978-0-12-804069-0.00002-2.

[173] J. Hook and H.E. Hall. *Solid state physics (2nd ed., The Manchester physics series),* Chichester ; New York: Wiley, 1991.

[174] H. Ibach and H. Lüth. *Solid-State Physics, An Introduction to Principles of Materials Science (2nd ed.),* Springer-Verlag, 1996.

[175] S.M. Sze and M.K. Lee. *Semiconductor devices physics and technology* (3rd ed.). Wiley, 2012.

[176] M. Winfried. *Surfaces of III–V and II–VI Compound Semiconductors in Semiconductor Surfaces and Interfaces,* Springer Berlin Heidelberg, pp. 84-116, 1993. https://doi.org/10.1007/978-3-662-02882-7_7

[177] C. Gardes, S. Bagumako, L. Desplanque, N. Wichmann, S. Bollaert, F. Danneville and Y.Roelens, '100 nm AlSb/InAs HEMT for Ultra-low-power Consimption, Low-noise Applications,' *The Scientific World Journal,* 6, 2014.

[178] M. Yuping Zeng *et al*, 'Quantum Well InAs/AlSb/GaSb Vertical Tunnel FET With HSQ Mechanical Support,' *IEEE Transactions on Nanotechnology,* 14(3), 580-584, 2015.

[179] K. Yoh, T. Moriuchi and M. Inoue. (1990). 'An InAs channel heterojunction field-effect transistor with high transconductance,' *IEEE Electron Device Letters,* 11(11), 526-528, 1990.

[180] A-M. Hoang, A. Dehzangi, S. Adhikary and M. Razeghi, 'High performance bias-selectable three colour short-wave/mid-wave/long-wave infrared photodetectors based on type-II InAs/GaSb/AlSb superlattices,' *Scientific Reports* 6(1):24144, 2016.

[181] E.H. Aifer, J.g> Tischler, J.H> Warner, I. Vurgaftman, W.W. Bewley, J.R. MEyer, J.C. Kim, L.J. Whitman, C.L. Canedy and E.M. Jackson, 'W-structures type=II superlattice long-wave infrared photodiodes with high quantum efficiency,' *Applied Physics Letters* 89(5):053519, 2006.

[182] O. Cathabard, R. Teisser, J. Devenson, J.C. Moreno and A.N. Baranox, 'Quantum cascade lasers emitting near 2.6 μm,' *Applied Physics Letters* 96(14):141110, 2010.

[183] Dorian Sanchez, Laurent Cerutti, and Eric Tournié. 'Single-mode monolithic GaSb vertical-cavity surface-emitting laser,' *Optics express* 20.14: 15540-15546, 2012.

[184] Lei, Wen, *et al*, 'GaSb: a new alternative substrate for epitaxial growth of HgCdTe,' *Journal of electronic materials* 43.8 (2014): 2788-2794, 2014.

[185] A. F. J. Levi, *Semiconductor band structure and heterostructures,* Essential Semiconductor Laser Device Physics, Morgan & Claypool Publishers, 2053-2571, 2018.

[186] T. Ihn, 'Chapter 5.1 Band engineering,' *Semiconductor Nanostructures Quantum States and Electronic Transport,* United States of America: Oxford University Press. p. 66, 2010.

[187] H. Kroemer, 'The 6.1 Åfamily (InAs, GaSb, AlSb) and its heterostructures: a selective review,' *Physica E: Low-dimensional Systems and Nanostructures,* 20(3-4):196-203, 2004.

[188] E. Lefebvre, M. Malmkvist, M. Borg, L. Desplanque, X. Wallart, G. Dambrine, S. Bollaert and J. Grahn, 'Gate-Recess Technology for InAs/AlSb HEMTs,' *IEEE Transactions on Electron Devices* 56(9) 1904-1911, 2009.

[189] G. Moschetti, *Ultra-low Power InAs/AlSb HEMTs for cryogenic low noise applications,* PhD thesis, Chamlers University of Technology, G}oteberg, Sweden (2012).

[190] K. Yoh, T. Moriuchi and M. Inoue, 'An InAs channel heterojunction field-effect transistor with high transconductance,' *IEEEElectron Device Letters* 11(11):526-528, 1990.

[191] D.Z. Ting, and Xavier Cartoixà. 'InAs/GaSb/AlSb Resonant Tunneling Spin Device Concepts,' *Physica E: Low-dimensional Systems and Nanostructures* 20, no. 3-4: 350-54, 2004.

[192] T. Shibata, J. Nakata, Y. Nanishi and M. Fujimoto. 'A Rutherford backscattering spectroscopic study of the aluminum antimonide oxidation process in air,' *Japanese journal of applied physics,* 33(4R), 1767, 1994.

[193] R. Mohammad, ‘The Electronic Band Structure of III (In, Al, Ga)-V (N, As, Sb) Compounds and Ternary Alloys,’ *Middle East Technical University*, 2005.

[194] T. Kruczek, K. A. Fedorova, G. S. Sokolovskii, R. Teissier, A. N. Baranov and E. U. Rafailov. ‘InAs/AlSb widely tunable external cavity quantum cascade laser around 3.2 μm,’ *Applied Physics Letters* 102, no. 1: 011124, 2013.

[195] T.B. Boykin, ‘Current-voltage calculations for InAs/AlSb resonant-tunneling diodes,’ *Physical Review B* 51.7: 4289, 1995.

[196] T. Nowozin, D. Bimberg, K. Daqrouq, M.N. Ajour and M. Awedh, ‘Materials for Future Quantum Dot-Based Memories,’ *J. Nanomater,* 2013, 1–6, 2013.

[197] I. Vurgaftman„ J. Meyer, and L. Ram-Mohan, ‘Band parameters for III–V compound semiconductors and their alloys,’ *Journal of applied physics* 89.11: 5815-5875, 2001.

[198] B.R. Nag, *Physics of Quantum Well Devices,* Solid-State Science and Technology Library, vol 7. Springer, Dordrecht, 2002.

[199] E.J. Koerperick, J.T. Olesberg, J.L. Hicks, J.P. Prineas and T.F. Boggess, ‘High-power MWIR cascaded InAs–GaSb superlattice LEDs,’ *IEEE Journal of Quantum Electronics,* 45(7), 849-853, 2009.

[200] Y. Wei *et al*, ‘Uncooled operation of type-II InAs GaSb superlattice photodiodes in the mid-wavelength infrared range,’ *Applied Physics Letters* 86.23: 233106, 2005.

[201] A.N. Baranov, N. Bertru, Y. Cuminal, G. Boissier, C. Alibert and A. Joullie, ‘Observation of room-temperature laser emission from type III InAs/GaSb multiple quantum well structures,’ *Applied physics letters,* 71(6), 735-737, 1997.

[202] B.M. Borg *et al*, ‘InAs/GaSb heterostructure nanowires for tunnel field-effect transistors,’ *Nano letters* 10.10: 4080-4085, 2010.

[203] M. Yokoyama, H. Yokoyama, M. Takenaka and S. Takagi, ‘Ultrathin body GaSb-on-insulator p-channel metal-oxide-semiconductor field-effect transistors on Si fabricated by direct wafer bonding,’ *Applied Physics Letters,* 106(7), 073503, 2015.

[204] M. Yokoyama, H. Yokoyama, M. Takenaka and S. Takagi, ‘InAs/GaSb-on-insulator single channel complementary metal-oxide-semiconductor transistors on Si structure,’ *Applied Physics Letters,* 109(21), 213505, 2016.

[205] H-Y. Chou *et al*, ‘Band offsets and trap-related electron transitions at interfaces of (100) InAs with atomic-layer deposited Al2O3,’ *Journal of Applied Physics* 120.23: 235701, 2016.

[206] R. Li *et al*, ‘InAs/AlGaSb heterojunction tunnel field-effect transistor with tunnelling in-line with the gate field,’ *physica status solidi* c 9.2: 389-392, 2012.

[207] L.ö Olsson, C.B. Andersson, M.C. Håkansson, J. Kanski, L. Ilver & U.O. Karlsson. ‘Charge accumulation at InAs surfaces,’ *Physical review letters,* 76(19), 3626, 1996.

[208] M. Saleem and Institute of Physics , publisher. *Quantum mechanics (IOP expanding physics),* Bristol [England] (Temple Circus, Temple Way, Bristol BS1 6HG, UK): IOP Publishing, 2015.

[209] A. Rae. *Quantum mechanics (Fifth ed.),* New York ; London, [England]: Taylor & Francis Group, 2007.

[210] B.J. Van Zehbroeck. ‘Density of States Calculation,’ [Available online: http://ecee.colorado.edu/ bart/book/dos.htm].

[211] S. Datta, *Electronic transport in mesoscopic systems (Cambridge studies in semiconductor physics and microelectronic engineering ; 3),* Cambridge ; New York: Cambridge University Press, 1995.

[212] E.R. Brown, C.D. Parker, L.J. Mahoney and K.M. Molvar, ‘Oscillations up to 712 GHz in InAs/AlSb diodes,’ *Society* 58:2291-2293, 1991.

[213] E. Ozbay, D.M. Bloom, D.H. CHow and J.N. Schulman, ‘1.7-ps microwave integrated-circuit-compatible InAs/AlSb resonant tunnelling diodes,’ *IEEE Electron Device Letters* 14(8):400-402, 1993.

[214] Y. C. Chou *et al.,* ‘Manufacturable and Reliable 0.1 µm AlSb/InAs HEMT MMIC Technology for Ultra-Low Power Applications,’ *2007 IEEE/MTT-S International Microwave Symposium,* pp. 461-464, 2007. doi: 10.1109/MWSYM.2007.380488.

[215] J.R Söderström, D.H. Chow and T.C. McGill. ‘New negative differential resistance device based on resonant interband tunneling,’ *Applied Physics Letters,* 55(11), 1094-1096, 1989.

[216] O. Tizno, A.R. Marshall, N. Fernández-Delgado, M. Herrera, S.I. Molina and M. Hayne. ‘Room-temperature operation of Low-voltage, Non-volatile, Compound-semiconductor Memory Cells,’ *Scientific reports,* 9(1), 1-8, 2019.

[217] M. Karalic, Tschirky, Wegscheider, Ensslin and Ihn. ‘Lateral p-n Junction in an Inverted InAs/GaSb Double Quantum Well,’ *Physical Review Letters,* 118(20), 206801, 2017.

[218] D. Lane and M. Hayne. ‘Simulations of Ultralow-Power Nonvolatile Cells for Random-Access Memory,’ *IEEE Transactions on Electron Devices,* 67(2), 474-480, 2020.

[219] J. Suñé, S. Lanzoni, R. Bez, P. Olivo and R. Riccò, ‘Transient simulation of the erase cycle of floating gate EEPROMs,’ *International Electron Devices Meeting 1991* [Technical Digest], Washington, DC, USA, pp. 905-908, 1991.

[220] 'AN10860 LPC313x NAND flash data and bad block management,' *NXP Semiconductors*, 2009.

[221] V. Virkkala, V. Havu, F. Tuomisto and M.J. Puska, 'Native point defects in GaSb' *Phys. Rev. B* 86 144101, 2012.

[222] L.A. Hanks, M. Hayne, A.R.J. Marshall and L. Ponomarenko, 'Transport of modulation-doped Al0. 2Ga0. 8Sb/GaSb heterojunctions,' *In Journal of Physics: Conference Series* Vol. 964, No. 1, p. 012006. IOP Publishing, 2018.

[223] U.K. Mishra and J. Singh, 'Structural Properties of Semiconductors,' *Semiconductor Device Physics and Design,* Springer, Dordrecht, 2008.

[224] S. Adachi and Tu, *Physical Properties of III-V Semiconductor CompoundsInP, InAs, GaAs, GaP, InGaAs and InGaAsP,* Physics Today, 47(2), 99-100, 1994.

[225] Y. Li, Zhang and Zeng, 'Electron mobility in modulation-doped AlSb/InAs quantum wells,' *J. of App. Phys.,* Vol.109(7), 2011.

[226] P. Drude, 'Zur Elektronentheorie der Metalle; II. Teil. Galvanomagnetische und thermomagnetische Effecte,' *Annalen der Physik,* 308 (11): 369–402, 1900. doi:10.1002/andp.19003081102

[227] S. Sze, Ng, Kwok Kwok, and ProQuest, *Physics of semiconductor devices (3rd ed.),* Hoboken, N.J.: Wiley-Interscience, 2007.

[228] Y.H. Kang and S. Hong, 'A simple Flash memory cell model for transient circuit simulation,' *Electron Device Letters,* IEEE, 26(8) pp. 563-565, 2005.

[229] '8th and 9th generation Intel Core processor families and Intel Xeon E processor families,' *Intel,* datasheet volume 1 of 2, revision 005, 2020.

[230] K. Prall, 'Scaling non-volatile memory below 30nm,' *IEEE NVSMW,* pp. 5-10, 2007.

[231] E.R. Brown, J.R. Söderstrom, C.D. Parker, L.J. Mahoney, K.M. Molvar and T.C. McGill, 'Oscillations up to 712 GHz in InAs/AlSb resonant-tunneling diodes,' *Applied Physics Letters,* 58(20):2291 - 2293, 1991.

[232] S. Wei and A. Zunger. 'Calculated natural band offsets of all II–VI and III–V semiconductors: Chemical trends and the role of cation d orbitals,' *Applied Physics Letters,* 72(16), 2011-2013, 1998.

[233] D. Lane, P. D. Hodgson, R. J. Potter, R. Beanland and M. Hayne, 'ULTRARAM: Toward the Development of a III–V Semiconductor, Nonvolatile, Random Access Memory,' *IEEE Transactions on Electron Devices,* vol. 68, no. 5, pp. 2271-2274, 2021. doi: 10.1109/TED.2021.3064788.

[234] B. Jacob, S. Ng and D. Wang, *Memory Systems: Cache, DRAM, Disk,* Elsevier Science, 2007.

[235] Kim, Yoongu, *et al*. 'Flipping bits in memory without accessing them: An experimental study of DRAM disturbance errors,' *ACM SIGARCH Computer Architecture News* 42.3: 361-372, 2014.

[236] J. Acharya, J. Wilt, B. Liu and J. Wu, 'Probing the Dielectric Properties of Ultrathin Al/Al2O3/Al Trilayers Fabricated Using in Situ Sputtering and Atomic Layer Deposition' *ACS Appl. Mater. Interfaces,* 10, 3, 3112–3120, 2018.

[237] Y. Jhan *et al*, 'Low-Temperature Microwave Annealing for Tunnel Field-Effect Transistor,' *IEEE Electron Device Letters,* vol. 36, no. 2, pp. 105-107, 2015. doi: 10.1109/LED.2014.2386213.

[238] M. Visciarelli, E. Gnani, A. Gnudi, S. Reggiani and G. Baccarani, enquoteDesign guidelines for GaSb/InAs TFET exploiting strain and device size, *Solid-State Electronics,* Vol. 129, pp 157-162, 2017. https://doi.org/10.1016/j.sse.2016.11.011.

[239] G. Zhou, R. Li, T. Vasen, M. Qi, S. Chae, Y. Lu and M. Wistey, 'Novel gate-recessed vertical InAs/GaSb TFETs with record high ION of 180 µA/µm at VDS= 0.5 V,' *2012 IEEE International Electron Devices Meeting,* pp. 32-6, 2012.

[240] S. Huang, G. Balakrishnan and D.L. Huffaker, 'Interfacial misfit array formation for GaSb growth on GaAs,' *Journal of Applied Physics,* 105(10), Journal of Applied Physics, 15 May 2009, Vol.105(10), 2009.

[241] S. Huang, G. Balakrishnan and D. L. Huffaker, 'Interfacial misfit array formation for GaSb growth on GaAs,' *Journal of Applied Physics* 105(10):0-5, 2009.

[242] A.P, Craig, P.J. Carrington, H. Liu and A.R.J Marshall, 'Characterisation of 6.1 ÅIII-V materials grown on GaAs and Si: A comparison of GaSb/GaAs epitaxy and GaSb/AlSb Si epitaxy,' *Journal of Crystal Growth,* 425:56-61, 2016.

[243] W. Li, S. Laaksonen, J. Haapamaa and M. Pessa. 'Growth of device-quality GaAs layer directly on (001) Ge substrates by both solid-source and gas-source MBE,' *Journal Of Crystal Growth,* 227, 104-107, 2001.

[244] E.P. Delli, J.J. Hayton, P.D. Carrington, V.R. Letka, P. Hodgson, E. Repiso, J.P Hayton, A.P. Craig, Q. Lu, R. Beanland, A. Krier and A.R.J. Marshall, 'Mid-Infrared InAs/InAsSb Superlattice nBn Photodetector Monolithically Integrated onto Silicon,' *ACS Photonics,* 6(2), 538-544, 2019.

[245] E. Delli, P. Hodgson, M. Bentley, E. Repiso-Menendez, A. Craig, Q. Lu, R. Beanland, A. Marshall, A. Krier and P. Carrington, 'Mid-infrared Type-II InAs/InAsSb Quantum

Wells Integrated on Silicon,' *Applied Physics Letters,* vol. 117, no. 13, 131103, 2020. https://doi.org/10.1063/5.0022235

[246] Overview of AutoCAD design software. [Available online: https://www.autodesk.co.uk/products/autocad/overview].

[247] KLayout mask design software. [Available online: https://www.klayout.de/].

[248] O. Tizno, *Design, fabrication and characterisation of a novel memory device based on III-V semiconductors,* Lancaster University, 2018. https://doi.org/10.17635/lancaster/thesis/486

[249] M.S. Bakir and J.D. Meindl. *Integrated Interconnect Technologies for 3D Nanoelectronic Systems,* Artech House Integrated Microsystems Series. Boston, Mass. ; London: Artech House, 2009.

[250] J. Sun and J. Kosel, 'Room temperature inductively coupled plasma etching of InAs/InSb in BCl3/Cl2/Ar,' *Microelectronic Engineering,* Vol. 98, pp. 222-225, 2012. https://doi.org/10.1016/j.mee.2012.07.018.

[251] L. Zhuang, L.F. Lester, R.J. Shul, C.G. Willison and R.P. Leavitt, 'Inductively-coupled plasma etching of III-V antimonides in BCl3/Ar and Cl2/Ar,' *Journal of Vacuum Science & Technology B: Microelectronics and Nanometer Structures Processing, Measurement, and Phenomena* 17, 965, 1999.

[252] Xue-Yang, *et al*, 'The Etching Properties of Al2O3 Thin Films in BCl3/Cl2/Ar Plasma,' *Ferroelectrics,* vol. 384, no. 1, pp. 39–46, 2009.

[253] X. Yang, Jong-Chang Woo, Doo-Seung Um and Chang-Il Kim. 'Dry Etching of Al2O3 Thin Films in O2/BCl3/Ar Inductively Coupled Plasma,' *Transactions on Electrical and Electronic Materials,* 11. 10.4313/TEEM.2010.11.5.202., 2010.

[254] Stan Augarten. 'The Chip Collection:State of the Art', 1983.

[255] K. Kajiyama, Y. Mizushima, and S. Sakata, 'Schottky barrier height of n-InxGa1-xAs diodes,' *Appl. Phys. Lett,* 23, 458, 1973. https://doi.org/10.1063/1.1654957

[256] S. Dutta, H. L. Bhat, and V. Kumar, 'The physics and technology of gallium antimonide: An emerging optoelectronic material,' *J. Appl. Phys,* 81, 5821, 1997.

[257] W. Rapson, 'The bonding of gold and gold alloys to non-metallic materials,' *Gold Bulletin,* 12(3), 108-114, 1979.

[258] J. Hölzl and Franz K. Schulte. 'Work function of metals,' *Solid surface physics,* Springer, Berlin, Heidelberg, 1-150, 1979.

[259] P. Oviroh, R. Akbarzadeh, D. Pan, R. Coetzee and T. Jen, ‘New development of atomic layer deposition: Processes, methods and applications,’ *Science and Technology of Advanced Materials,* 20(1), 465-496, 2019.

[260] M. Milojevic, F.S. Aguirre-Tostado, C.L. Hinkle, H.C. Kim, E.M. Vogel, J. Kim and R.M. Wallace, ‘Half-cycle atomic layer deposition reaction studies of Al2O3 on In0.2Ga0.8As (100) surfaces,’ *Appl. Phys. Lett.,* 93, 202902, 2008.

[261] L. Zhou, B. Bo, X. Yan, C. Wang, Y. Chi and X. Yang, ‘Brief Review of Surface Passivation on III-V Semiconductor,’ *Crystals,* 8(5), Crystals, 2018 May, Vol.8(5), 2018.

[262] N. Li, Harmon, Hyland, Salzman, Ma, Xuan, and Ye, ‘Properties of InAs Metal-oxide-semiconductor Structures with Atomic-layer-deposited Al 2 O 3 Dielectric,’ *Applied Physics Letters* 92, no. 14: Vol.92(14), 2008.

[263] G. D’Acunto *et al*, ‘Atomic Layer Deposition of Hafnium Oxide on InAs: Insight from Time-Resolved in Situ Studies,’ *ACS Applied Electronic Materials* 2, 12, 3915–3922, 2020.

[264] J. Henrie, S. Kellis, S.M. Schultz and A. Hawkins, ‘Electronic color charts for dielectric films on silicon,’ *Opt. Express 12,* 1464-1469, 2004.

[265] SigmaAldrich Buffered oxide etchant (BOE) 10:1, [Available online: https://www.sigmaaldrich.com/catalog/product/aldrich/901621].

[266] Dongchul Suh, ‘Etch Characteristics and Morphology of Al2O3/TiO2 Stacks for Silicon Surface Passivation,’ *Sustainability* 11, no. 14, 2019.

[267] M.M. Winterkorn, A. Dadlani, T. Kim, T.S. English, K.L. Harrison, J. Provine and F.B. Prinz, ‘ATOMIC LAYER DEPOSITED ETCH STOP LAYERS FOR HYDROFLUORIC ACID,’ *Conference: 2016 Solid-State, Actuators, and Microsystems Workshop,* 133-136, 2016.

[268] K. Williams, K. Gupta and M. Wasilik, ‘Etch Rates for Micromachining Processing—Part II,’ *Journal of Microelectromechanical Systems,* 12. 761 - 778, 2004.

[269] H.C. Lin, G. Brammertz, K. Martens, G. de Valicourt, L. Negre, W.E. Wang and M. Heyns, ‘The Fermi-level efficiency method and its applications on high interface trap density oxide-semiconductor interfaces,’ *Applied Physics Letters,* 94(15), 153508, 2009.

[270] M. R. Zakaria and M.N. Hashim, ‘An overview and simulation study of conventional flash memory floating gate device using concept FN tunnelling mechanism,’ *Proc. of V-th Int. Conf. on Intelligent Systems, Modeling and Simulation,* pp. 775-780, 2014.

[271] R. Rajput and R. Vaid, ‘Flash memory devices with metal floating gate/metal nanocrystals as the charge storage layer: A status review,’ *Facta universitatis - series: Electronics and Energetics,* 33, 155-167, 2020. DOI:10.2298/FUEE2002155R.

[272] J.G. Zhu, X.L. Yang and M. Tao, ‘Low-resistance titanium/n-type silicon (100) contacts by monolayer selenium passivation,’ *Journal of Physics D: Applied Physics* 40, 547, 2007.

[273] D. Lane and M. Hayne, ‘Simulations of resonant tunnelling through InAs/AlSb heterostructures for ULTRARAM memory,’ *Journal of Physics D: Applied Physics* 54, 355104, 2021.

[274] P. Harrison P and A. Valavanis *Quantum wells, wires and dots: Theoretical and computational physics of semiconductor nanostructures* (Wiley, England). 4th ed. p. 163, 2016.

[275] D. Lane, P.D. Hodgson, R.J. Potter, R. Beanland and M. Hayne, ‘Demonstration of a Fast, Low-voltage, III-V Semiconductor, Non-volatile Memory,’ *2021 5th IEEE Electron Devices Technology & Manufacturing Conference* (EDTM), Chengdu, 2021; https://doi.org/10.1109/EDTM50988.2021.9420825

[276] P.D. Hodgson *et al*, ‘III-V non-volatile ULTRARAM™ memory on Si,’ *2021 APL Materials: Materials Challenges for Memory* (MCfM), virtual conference, 2021.

[277] H.P. Hwang *et al*, ‘High peak-to-valley current ratio In/sub 0.3/Ga/sub 0.7/As/In/sub 0.29/Al/sub 0.71/As resonant tunneling diodes grown on GaAs,’ *1994 International Electron Devices and Materials Symposium* (EDMS), Hsinchu, Taiwan pp. 3-3, 1994.

[278] T. Kubis, *Quantum transport in semiconductor nanostructures,* PhD thesis, Technische Universität München, 2009.

[279] M.Feiginov, ‘Frequency Limitations of Resonant-Tunnelling Diodes in Sub-THz and THz Oscillators and Detector,’ *J. Infrared Milli Terahz Waves* 40, 365, 2019. https://doi.org/10.1007/s10762-019-00573-5

[280] Stefan Birner, *Modeling of semiconductor nanostructures and semiconductor-electrolyte interfaces,* Muenchen Technische Univ., Germany, 2011.

[281] G. Arfken, *Mathematical Methods for Physicists,* 3rd ed. Orlando, FL: Academic Press, pp. 963-964, 1985.

[282] A.J. Bestwick, *Quantum Edge Transport in Topological Insulators* (Thesis). Stanford University, 2015.

[283] D.J. Griffiths, *Introduction to Quantum Mechanics (2nd ed.),* Prentice Hall, 2004.

[284] T. Nowozin, A. Marent, L. Bonato, A. Schliwa, D. Bimberg, E. Smakman and M. Hayne, ‘Linking structural and electronic properties of high-purity self-assembled GaSb/GaAs quantum dots,’ *Physical Review B,* 86(3), 035305, 2012.

[285] D. Ahmad, I. Boogaert, J. Miller, R. Presswell and H. Jouhara, ‘Hydrophilic and hydrophobic materials and their applications,’ *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects,* 40:22, 2686-2725, 2018.

[286] Microchemicals Resist Adhesion - MicroChemicals. [Available online: https://www.microchemicals.com/technical_information/resist_adhesion.pdf].

[287] Allresist Gmbh. ‘FAQs concerning photoresists from Allresist,’ [Available online: https://www.allresist.com/wp-content/uploads/sites/2/2016/03/faqs_photoresists_eng_2016.pdf].

[288] W. Ng, Y. Lu, H. Liu, C. Carmalt, I. Parkin and A. Kenyon, ‘Controlling and modelling the wetting properties of III-V semiconductor surfaces using re-entrant nanostructures,’ *Scientific Reports,* 8, 2018.

[289] Y. Zhao, J.Li, J. Hu, L. Shu and X. Shi, ‘Fabrication of Super-Hydrophobic Surfaces with Long-Term Stability,’ *Journal of Dispersion Science and Technology,* 32:7, 969-974, 2011.

[290] G.C. Desalvo, ‘Citric Acid Etching of GaAs[sub 1-x]Sb[sub x], Al[sub 0.5]Ga[sub 0.5]Sb, and InAs for Heterostructure Device Fabrication,’ *Journal of The Electrochemical Society,* 141(12), pp.3526-3531, 1994.

[291] C. Gatzke, S. Webb, K. Fobelets & R. Stradling, ‘In-situ monitoring of the selective etching of antimonides in GaSb/AlSb/InAs heterostructures using Raman spectroscopy,’ *Compound Semiconductors 1997. Proceedings of the IEEE Twenty-Fourth International Symposium on Compound Semiconductors,* 337-340, 1997.

[292] O. Klin, N. Snapi, Y. Cohen and E. Weiss, ‘A study of MBE growth-related defects in InAs/GaSb type-II supperlattices for long wavelength infrared detectors,’ *Journal of Crystal Growth,* Vol. 425, pp. 54-59, 2015.

Appendix A

nextnano Simulation Calculation Details

A.1 Poisson-Schrödinger

The nextnano++ software package provides nanostructure simulations calculated from Schrödinger-Poisson solutions. nextnano++ is a console program written in C++, whose computational scheme is as follows:

- The input file is processed: material parameters obtained from the database is mapped to the specified simulation mesh (1-D in this work).
- Strain is calculated and band-edges are positioned accordingly
- The Schrödinger equation and Poisson equation are solved self consistently, including:

– Built-in potential

– Potentials shifted according to applied bias on contacts

– Quasi-Fermi levels calculated from drift-diffusion and current continuity equations (wave functions and potential fixed)

– Carrier densities and potentials are calculated from Schrödinger-Poisson equations

A.1.1 Strain

In this work, we employ the strain minimisation model for all nextnano calculations. This is calculated prior to and independently of the Schrödinger, Poisson and drift-diffusion equations. The strain tensor is calculated numerically by minimising elastic energy [280].

A.1.2 Schrödinger equation

The Schrödinger equation is a partial differential equation (PDE) which describes how the wavefunction of a physical system evolves, and therefore can be used to compute the energy states of a given quantum system. It is given by

$$i\hbar\frac{\partial}{\partial t}\psi(\vec{r},t) = H\psi(\vec{r},t)\ , \tag{A.1}$$

where H is the Hamiltonian operator which characterises the energy of the system. Thus, it is the sum of the system's kinetic (K) and potential (V) energy: $H \equiv K + V$. For non-relativistic particles, the Hamiltonian for charged particles in a semiconductor can be expressed as

$$H = -\frac{\hbar^2}{2m^*}\nabla^2 + V(\vec{r},t)\ , \tag{A.2}$$

where m^* is the effective mass.

Exact solutions for a PDE are not always obtainable and instead a solution is often approximated by a numerical method. Such techniques split the domain of the function we wish to solve into a discrete set of function values which approximate the original. The partitioning into subdomains is called a mesh or grid. The nextnano++ software implements the finite difference method (FDM) to produce numerical solutions to the Schrödinger equation. The mathematics of this technique can be summarised as follows: The derivative of a function ($f(x)$) is defined as:

$$\frac{\partial f(x)}{\partial x} = \lim_{h\to 0}\frac{f(x+h)-f(x)}{h}\ . \tag{A.3}$$

Consequently, for a small h, we have

$$\frac{\partial f}{\partial x} \approx \frac{f(x+h)-f(x)}{h}\ .$$

If lines on the grid are separated by h, this approximation can be used to solve the PDE at each point, where a finer mesh increases the accuracy (smaller h).

A.1.3 Poisson equation

The Poisson equation is an electrostatics equation describing a potential field from a charge distribution (classical):

$$\nabla\cdot(\epsilon_0\epsilon_r\nabla\psi) = -\rho\ , \tag{A.4}$$

where ϵ_0 is vacuum permittivity, ϵ_r is the dielectric constant, ψ is the electrostatic potential and ρ is the charge density distribution. This charge density distribution is the semiconductor material is given by:

$$\rho = e(-n + p + N_d^+ - N_a^-)\ , \tag{A.5}$$

where n is the electron density, p is the hole density and N_d^+ and N_a+ are ionised donor and acceptor concentrations respectively. Within nextnano, the Poisson equation is solved via numerical iterations of Newton's method [281]. The software handles the boundary conditions: for non-equilibrium situations such as applied potential across the system, built-in (equilibrium) potential is first calculated using Neumann boundaries ($\partial\psi/\partial x = 0$) before Dirichlet boundary conditions are implemented to alter the field at the contacts, accounting for the built-in potential.

A.1.4 Drift diffusion

The quasi-Fermi levels (*i.e.* Fermi-levels at non-equilibrium), $E_{F,n/p}$ and carrier densities, n or p are linked by the drift-diffusion equation which is a classical description of carrier movement in an electric field

$$\begin{aligned} \vec{J_n} &= en\mu_n\nabla\psi + \mu_n k_B T\nabla n = e\mu_n n\nabla E_{F,n} \\ \vec{J_p} &= ep\mu_p\nabla\psi - \mu_p k_B T\nabla p = e\mu_p p\nabla E_{F,p}\ , \end{aligned} \tag{A.6}$$

where $\vec{J}$ is current density and µ is carrier mobility. The ∇n signifies the diffusion of carriers, whilst the $\nabla\phi$ term signifies carrier drift in the field. If one assumes that the carrier recombination and generation is negligible, current continuity for carriers in the semiconductor simplifies to

$$\nabla\dot{J} = 0\ , \tag{A.7}$$

which can be used with equation A.6 to calculate the quasi-Fermi levels which, in turn, are inserted into the Poisson equation to be self-consistently solved with the Schrödinger equation.

It is clear that the drift-diffusion and current continuity formulation does not describe quantum-transport mechanisms such as resonant tunnelling or band-to-band tunnelling. As such, the next-nano++ software is not capable of fully describing the operation of ULTRA**RAM**™ P/E cycles or the NORMALLY-OFF conductivity-modulation mechanism.

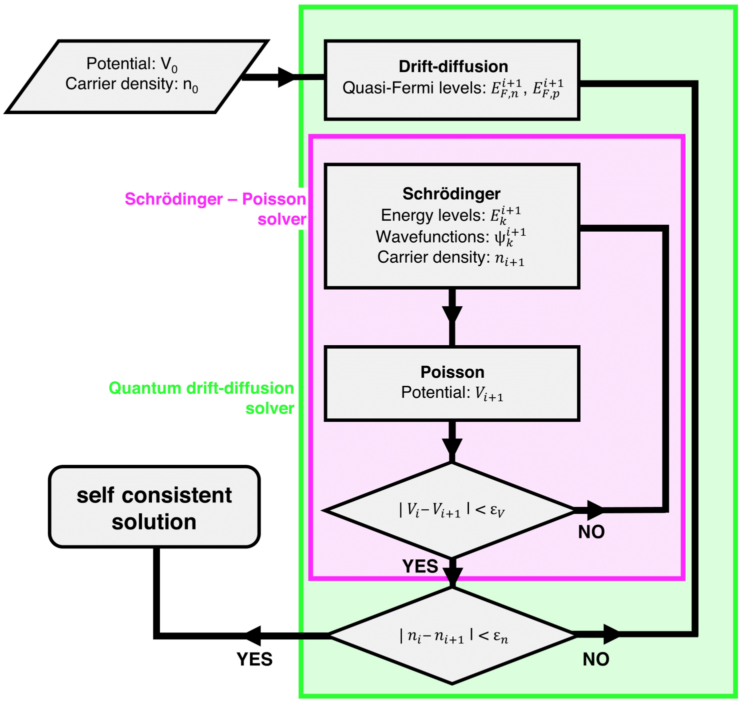
A.1.5 Self-consistency

Self-consistent solutions of the Schrödinger and Poisson equations are achieved as both have inter-dependent properties: the potential ($V(\vec{r})$) and carrier density ($n(\vec{r})$). Self-consistent solving involves using the output from one equation as the input of the other in an iterative fashion, until the difference between the solutions reduces below a given limit (user-defined).

This process is demonstrated more clearly by the flowchart presented in Fig. A.1. A trial potential, V_0, is used to solve the Schrödinger equation. Then, the $n(\vec{r})$ is determined and used in the

Poisson equation to obtain a new potential value, V_1, which is then fed back into the Schrödinger equation. This proceeds until he difference between them is $< \epsilon_V$, which is a user-defined limit.

Before this process begins, the quasi-Fermi levels are determined from the carrier-continuity requirement and are added to the Schrödinger-Poisson iteration. The quasi-Fermi level calculations are refreshed with each iteration forming a second self-consistency loop which requires similar carrier density output from the Schrödinger equation and Poisson equation before returning a solution.



Potential: V_0
Carrier density: n_0
Drift-diffusion
Quasi-Fermi levels: $E_{F,n}^{i+1}$, $E_{F,p}^{i+1}$
Schrödinger – Poisson solver
Schrödinger
Energy levels: E_k^{i+1}
Wavefunctions: ψ_k^{i+1}
Carrier density: n_{i+1}
Quantum drift-diffusion solver
Poisson
Potential: V_{i+1}
self consistent solution
$|V_i - V_{i+1}| < \varepsilon_V$
NO
YES
$|n_i - n_{i+1}| < \varepsilon_n$
YES
NO

Figure A.1: Flow chart of the Schrödinger-Poisson solving technique of nextnano++. Two self-consistency loops are formed by the requirement of similar results for potential (V) and carrier density (n) from the Schrödinger equation and Poisson equation.

A.2 Non-equilibrium Green's Functions

A.2.1 Introduction

This appendix has been included to provide an introduction to the NEGF framework for those who may use this technique as a simulation tool for nanoscale devices [211]. It serves as a basic entry point for the non-theoretician and is by no means a complete description of the NEGF technique used within the nextnano.MSB software package, which is detailed in [167].

First, we consider how we view a current flow in small devices. This is typically viewed with source (S) and drain (D) terminals connected by a channel through which electrons flow [Fig. A.2**(a)**]. Within the channel, there are electronic states available for electron transport. Here, in describing transport, we need to include both mechanics and entropy-driven processes. If we begin with the Schrödinger equation;

$$E\{\psi\} = [\mathrm{H}]\{\psi\}\ , \tag{A.8}$$

we can write [H] in the form of matrices which describe the channel. This describes the isolated channel and not transport [Fig. A.2**(a)**]. We also consider a small range of energies (dE) in the channel, with number of states, D_0, then

$$D_0 = dE\ D(E)\ , \tag{A.9}$$

where $D(E)$ is the DOS. We also have a number of electrons, N, in the channel [Fig. A.2**(a)**].

In order to formulate current flow from the contacts through the channel, we consider that there are N electrons in the channel which can leave the first contact at a rate given as ν_1 and can leave the second contact at a rate of ν_2 [Fig. A.2**(a)**]. Simultaneously, new electrons are received into the channel from the contact. This depends on the DOS for the channel where s_1 and s_2 describe the incoming rate of electrons from the first and second contact respectively.

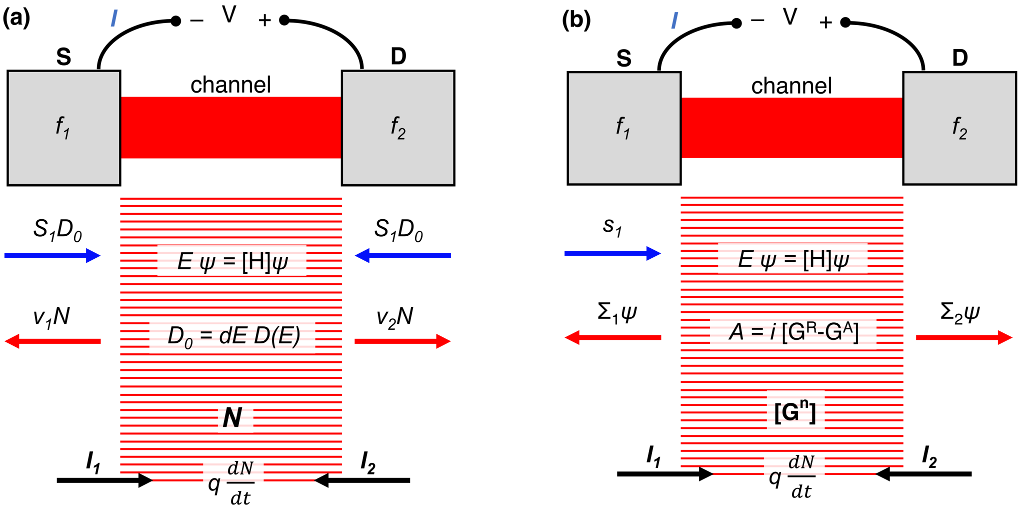


Figure A.2: Schematic representation of a simple channel to introduce the NEGF equations. **(a)** Semi-classical depiction for reference. **(b)** Quantum (Schrödinger) equivalent.

In general, we can use this formulation and write

$$\begin{aligned} \frac{dN}{dt} &= -(\nu_1 + \nu_2)N \leftarrow \textbf{outflow} \\ &\quad + (s_1 + s_2)D_0 \leftarrow \textbf{inflow} \; . \end{aligned} \tag{A.10}$$

The Schrödinger equation can now be suitably modified to include the inflow and outflow between the contacts and channel [Fig. A.2**(b)**]. The number of electrons in the channel is related to the electron probability as there are many, *i.e.*

$$N = \tilde{\psi}^\dagger \tilde{\psi} = \psi^\dagger \psi \; , \tag{A.11}$$

where $\tilde{\psi}$ relates to the time-dependent Schrödinger equation, as current flow will require time-dependent considerations. We now write the problem as

$$\begin{aligned} E\{\psi\} &= [\mathrm{H}]\{\psi\} \\ &\quad + [\Sigma_1 + \Sigma_2]\{\psi\} \leftarrow \textbf{outflow} \\ &\quad + \{s_1\} \leftarrow \textbf{inflow} \; , \end{aligned} \tag{A.12}$$

where Σ_1, Σ_2 and s_1 are determined by boundary conditions . Only one inflow (s_1) is included in this picture as otherwise the solutions give ψ in terms of s_1 and s_2, *i.e.*

$$\psi \sim (s_1 + s_2) \; . \tag{A.13}$$

From this, $\psi^\dagger \psi$ is calculated as

$$\begin{aligned} \psi^\dagger \psi &\sim (s_1 + s_2)^\dagger (s_1 + s_2) \\ &= s_1^\dagger s_1 + s_2^\dagger s_2 + s_1^\dagger s_2 + s_2^\dagger s_1 \; , \end{aligned} \tag{A.14}$$

yielding a cross-product of left and right contacts. This represents an interference between the contacts which is unphysical for electron transport.

We require a theory which works directly with wavefunction products such as $\psi^\dagger \psi$ rather than ψ in order to avoid interference terms. The NEGF equations are a set of formulae which apply to wavefunction products. The first of these is

$$\psi \psi^\dagger = [\mathrm{G}^n] \; , \tag{A.15}$$

which is an $n \times n$ matrix for a column vector ψ of n entries. The trace of [G^n] gives $\psi^\dagger \psi$. One can consider [G^n] as a matrix equivalent of N, representing the number of electrons in the channel.

We now define

$$\{s s^\dagger\} = [\Sigma^{in}] \; , \tag{A.16}$$

to describe the strength of the source. Note at this point that we cannot superimpose wavefunctions when we have multiple electrons, but we can superimpose their $\psi\psi^\dagger$, which motivates this formulation.

Returning to the Schrödinger equation, we define a new function, Σ, as

$$\Sigma = \Sigma_1 + \Sigma_2 \, , \tag{A.17}$$

then the Schrödinger equation can be written as

$$E\psi = H\psi + \Sigma\psi + s \, , \tag{A.18}$$

which can be rearranged to

$$[EI - H - \Sigma]\psi = s \, , \tag{A.19}$$

where I is an identity matrix of similar dimensions to [H] and [Σ]. We can calculate ψ from the above using an inverse:

$$\psi = \mathrm{G}^R\{s\} \, , \tag{A.20}$$

where

$$\mathrm{G}^R = [EI - H - \Sigma]^{-1} \, , \tag{A.21}$$

which is known as the retarded Green's function. Similarly, $\psi^\dagger$ can be calculated as

$$\psi^\dagger = \{s\}^\dagger \mathrm{G}^A \, , \tag{A.22}$$

where

$$\mathrm{G}^A = [\mathrm{G}^R]^\dagger \, , \tag{A.23}$$

which is known as the advanced Green's function. From this we can proceed to calculate $\psi\psi^\dagger$

$$\underbrace{\psi\psi^\dagger}_{\mathrm{G}^n} = \mathrm{G}^R \underbrace{ss^\dagger}_{\Sigma^{in}} \mathrm{G}^A \, . \tag{A.24}$$

This is often rewritten with the introduction of new functions $\mathrm{G}^<$ and $\Sigma^<$ where

$$-i\mathrm{G}^< \equiv \mathrm{G}^n \, , \tag{A.25}$$

$$-i\Sigma^< \equiv \Sigma^{in} \, , \tag{A.26}$$

Thus, we can write

$$\mathrm{G}^n = \mathrm{G}^R\Sigma^{in}\mathrm{G}^A \tag{A.27}$$

and

$$\mathrm{G}^< = \mathrm{G}^R\Sigma^<\mathrm{G}^A \, , \tag{A.28}$$

which completes the basic set of NEGF equations [211].

Semi-classical model

From the previous formulation of the transport system [Fig. A.2**(a)**], we can consider current flowing in two directions: left to right (I_1) and right to left (I_2). Considering the inflow out outflow of electrons:

$$\begin{aligned}\frac{dN}{dt} &= S_1 D_0 - \nu_1 N \ \leftarrow \sim I_1 \\ &+ S_2 D_0 - \nu_2 N \ \leftarrow \sim I_2\end{aligned} \tag{A.29}$$

where we consider current as electrons per second (*i.e.* charge is not yet included for coulombs per second). At steady state,

$$\frac{dN}{dt} = 0 \tag{A.30}$$

$$\therefore \frac{N}{D_0} = \frac{S_1 + S_2}{\nu_1 + \nu_2} . \tag{A.31}$$

The steady state current can therefore be determined from

$$I_1 = D_0 (S_1 - \nu_1 \frac{N}{D_0}) \tag{A.32}$$

$$\implies I_1 = D_0 \frac{\nu_2 S_1 - \nu_1 S_2}{\nu_1 + \nu_2} = -I_2 . \tag{A.33}$$

We now consider the system if only a single contact is connected. From statistical mechanics, we expect the system to reach equilibrium and the channel to be described by the same Fermi function (f) of the contact, *i.e.* for the first contact

$$\frac{N}{D_0} = \frac{S_1}{\nu_1} = f_1(E) \implies S_1 = \nu_1 f_1(E) . \tag{A.34}$$

Similarly, for the second contact

$$S_2 = \nu_2 f_2(E) . \tag{A.35}$$

Substituting into equation A.33:

$$I_1 = -I_2 = D_0 \frac{\nu_1 \nu_2}{\nu_1 + \nu_2} (f_1 - f_2) , \tag{A.36}$$

so current depends on the difference between the Fermi functions between contacts and the rate of electrons through the interfaces. Here, this current represents a small energy range dE. The energies conduct independently such that we can write the total currents as

$$I_1^T = -I_2^T = \int_{-\infty}^{\infty} dE\ D(E) \frac{\nu_1 \nu_2}{\nu_1 + \nu_2}(f_1 - f_2)\ , \tag{A.37}$$

for a complete current expression. Similarly, N can be written as

$$N = \int_{-\infty}^{\infty} dE\ D(E) \frac{\nu_1 f_1 + \nu_2 f_2}{\nu_1 + \nu_2}\ , \tag{A.38}$$

following the steady-state expression.

The purpose of this semi-classical approach is to provide a physical picture to understand the formulation of the quantum model which we will now outline.

A.2.2 Quantum model

To see the inflow and outflow in the quantum model we begin with the time-dependent Schrödinger equation

$$i\hbar \frac{d}{dt}\{\tilde{\psi}\} = [\mathrm{H}]\{\tilde{\psi}\} \tag{A.39}$$

$$\{\tilde{\psi}\} = \{\psi\} e^{\frac{-iEt}{\hbar}}\ . \tag{A.40}$$

We apply boundary conditions at the contact interfaces, which gives rise to the terms $\{s_1\}$ for inflow and Σ_1, Σ_2 for outflow [Fig. A.2**(b)**].

As $N = \tilde{\psi}^\dagger \tilde{\psi}$, and the time-dependent Schrödinger equation gives us $d\tilde{\psi}/dt$, we can employ the chain rule:

$$i\hbar \frac{d}{dt}(\tilde{\psi}^\dagger \tilde{\psi}) = i\hbar \left(\frac{d}{dt}\tilde{\psi}^\dagger\right) \tilde{\psi} + \tilde{\psi} \left(i\hbar \frac{d}{dt}\tilde{\psi}\right)\ . \tag{A.41}$$

We can now substitute the terms of equation A.18 to calculate this. For the first term, $[\mathrm{H}]\tilde{\psi}$ only:

$$\begin{aligned} i\hbar \left(\frac{d}{dt}\tilde{\psi}^\dagger\right) \tilde{\psi} + \tilde{\psi} \left(i\hbar \frac{d}{dt}\tilde{\psi}\right) &= -(H\tilde{\psi})^\dagger + \tilde{\psi}^\dagger (H\tilde{\psi}) \\ &= -\tilde{\psi}^\dagger H^\dagger \tilde{\psi} + \tilde{\psi}^\dagger H \tilde{\psi} \\ &= \tilde{\psi}^\dagger (H - H^\dagger)\tilde{\psi}\ . \end{aligned} \tag{A.42}$$

The Hamiltonian must be hermitian, so the H term is zero. This is as expected as there should be no dN/dt for an isolated channel. For the second term (outflow, equation A.18):

$$i\hbar \frac{d}{dt}(\tilde{\psi}^\dagger \tilde{\psi}) = \tilde{\psi}^\dagger (\Sigma - \Sigma^\dagger)\tilde{\psi}\ , \tag{A.43}$$

where

$$\Sigma = \Sigma_1 + \Sigma_2\ , \tag{A.44}$$

as Σ is non-hermitian, the term is non-zero. Lastly, the inclusion of the inflow term (equation A.18):

$$i\hbar\frac{d}{dt}(\tilde{\psi}^{\dagger}\tilde{\psi}) = -\tilde{s}^{\dagger}\tilde{\psi} + \tilde{\psi}^{\dagger}\tilde{s}\ . \tag{A.45}$$

We now introduce a new matrix, Γ:

$$\Gamma \equiv i(\Sigma - \Sigma^{\dagger})\ , \tag{A.46}$$

which is like the anti-hermitian part of Σ (outflow). Combining $N = \tilde{\psi}^{\dagger}\tilde{\psi}$ with equation A.41 gives us

$$\frac{dN}{dt} = \tilde{\psi}^{\dagger}\frac{\Gamma_1 + \Gamma_2}{\hbar}\tilde{\psi} + \frac{1}{i\hbar}(\tilde{\psi}\tilde{s} - \tilde{s}^{\dagger}\tilde{\psi})\ . \tag{A.47}$$

Note that comparing with the semi-classical picture, Γ is comparable to ν and $\tilde{s}$ is comparable to S (Fig. A.3). For current flow (now including electron charge):

$$\frac{dN}{dt} = \frac{1}{q}(I_1 + I_2)\ . \tag{A.48}$$

By comparison with equation A.47 we now have

$$\begin{aligned} \frac{I_1 + I_2}{q} &= \tilde{\psi}^{\dagger}\frac{\Gamma_1 + \Gamma_2}{\hbar}\tilde{\psi} \leftarrow \textbf{outflow} \\ &+ \frac{1}{i\hbar}(\tilde{\psi}\tilde{s} - \tilde{s}^{\dagger}\tilde{\psi}) \leftarrow \textbf{inflow}\ , \end{aligned} \tag{A.49}$$

where we are now back to steady-state currents such that $\tilde{\psi} \to \psi$. Using the retarded and advanced Green's functions discussed previously we can replace ψ and $\psi^{\dagger}$ in the above equation for the inflow term which becomes

$$\frac{1}{i\hbar}(s_1^{\dagger}\mathrm{G}^{A}s_1 - s_1^{\dagger}\mathrm{G}^{R}s_1) = s_1^{\dagger}\frac{A}{\hbar}s_1\ , \tag{A.50}$$

where we define a quantity A, which acts like the density of states

$$A = i[\mathrm{G}^{R} - \mathrm{G}^{A}] \tag{A.51}$$

and we now add the second contact source term to complete the picture [Fig. A.3**(b)**] where the connection to the semi-classical model is clear:

$$\begin{aligned} \frac{I_1 + I_2}{q} &= -\psi^{\dagger}\frac{\Gamma_1}{\hbar}\psi - \psi^{\dagger}\frac{\Gamma_1}{\hbar}\psi \leftarrow \textbf{outflow} \\ &+ s_1^{\dagger}\frac{A}{\hbar}s_1 + s_2^{\dagger}\frac{A}{\hbar}s_2 \leftarrow \textbf{inflow}\ . \end{aligned} \tag{A.52}$$

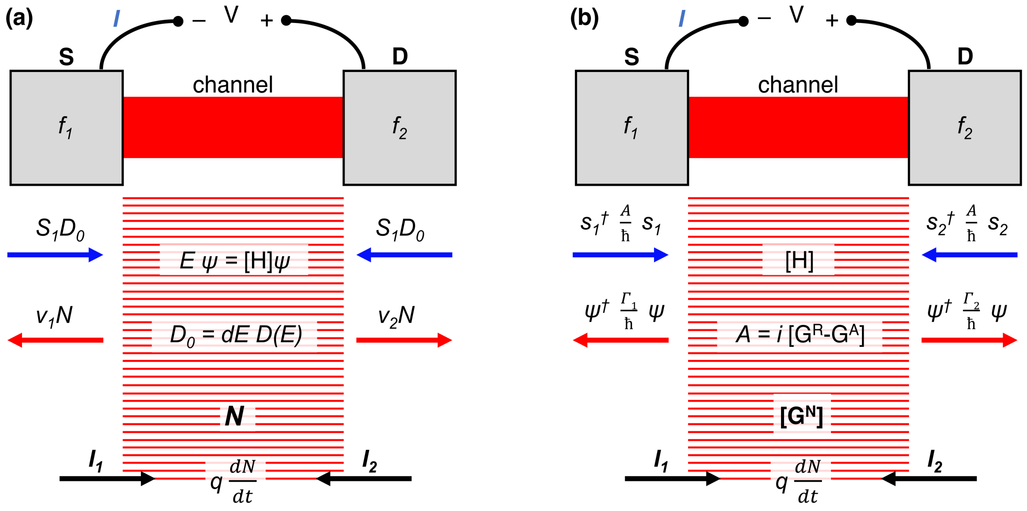


Figure A.3: Schematic representation of a simple channel for the NEGF formalism. **(a)** Semi-classical depiction for reference. **(b)** NEGF version using variables defined in the text.

A.2.3 NEGF equations

Next, we aim to take this quantum picture into the standard NEGF equations. Firstly, we take the matrix formed by $\psi\psi^\dagger$:

$$\psi\psi^\dagger = \frac{1}{2\pi}[\mathrm{G}^n] \,, \tag{A.53}$$

which is like the matrix form of N.

We also have a quantity related to the source term, s

$$ss^\dagger = \frac{1}{2\pi}[\Sigma^{in}] \,. \tag{A.54}$$

Returning to the current equations previously derived we have

$$\frac{I_1}{q} = -\psi^\dagger\frac{\Gamma_1}{\hbar}\psi + s_1^\dagger\frac{A}{\hbar}s_1 \,, \tag{A.55}$$

as each term gives just a number, they are equal to their trace, *i.e.*

$$I_1 = q\mathbf{Tr}\left[-\psi^\dagger\frac{\Gamma_1}{\hbar}\psi + s_1^\dagger\frac{A}{\hbar}s_1\right] \,, \tag{A.56}$$

writing as a trace allows us to move things around as long as we preserve cyclic order such that we can write

$$I_1 = q\mathbf{Tr}\left[\frac{-\Gamma_1}{\hbar}\psi\psi^\dagger + s_1s_1^\dagger\frac{A}{\hbar}\right] \,. \tag{A.57}$$

This equation is now expressed in a form containing $\psi\psi^\dagger$ and $s_1 s_1^\dagger$ which are matrices such that taking taking the trace is more meaningful. We can now substitute these for their Green's functions:

$$I_1 = \frac{q}{\hbar}\mathbf{Tr}\left[-\Gamma_1 \mathrm{G}^n + \Sigma_1^{in} A\right] \ , \tag{A.58}$$

G^n, A, Σ^{in} and Γ are analogous to their semi-classical counterparts as demonstrated in Figure A.3. For example the total number of electrons can be written as

$$\textbf{total number of electrons} = \int_{-\infty}^{\infty} dE \frac{\mathbf{Tr}[\mathrm{G}^n]}{2\pi} \ , \tag{A.59}$$

where the 2π appears from Parseval's theorem.

Similarly, the total number of states can be calculated as

$$\begin{aligned}\int_{-\infty}^{\infty} dE \frac{\mathbf{Tr}[A(E)]}{2\pi} =& \frac{i}{2\pi}\mathbf{Tr}\left[\int_{-\infty}^{\infty} dE[\mathrm{G}^R - \mathrm{G}^A]\right] \\ &= \textbf{total number of states} \ .\end{aligned} \tag{A.60}$$

In the semi-classical model, we argued that $S_1 = \nu_1 f_1(E)$ by considering a single contact. By analogy, for the quantum model we can write

$$\Sigma_1^{in}(E) = \Gamma_1(E) f_1(E) \tag{A.61}$$

$$\Sigma_2^{in}(E) = \Gamma_2(E) f_2(E) \ . \tag{A.62}$$

We now return to the NEGF equation described in the introduction

$$\mathrm{G}^n = \mathrm{G}^R \Sigma^{in} \mathrm{G}^A \ . \tag{A.63}$$

Consider at equilibrium where Fermi functions are equal, $f_1 = f_2 = f_0$, then

$$\begin{aligned}\Sigma^{in} = \Sigma_1^{in} + \Sigma_2^{in} &= \Gamma_1 f_1 + \Gamma_2 f_2 \\ &= (\Gamma_1 + \Gamma_2) f_0 \\ &= \Gamma f_0 \textbf{ (at equilibrium)} \ .\end{aligned} \tag{A.64}$$

Combing with equation A.63:

$$\mathrm{G}^n = [\mathrm{G}^R \Gamma \mathrm{G}^A] f_0 \ . \tag{A.65}$$

As previously shown,

$$
\begin{aligned}
[\mathrm{G}^R]^{-1} &= EI - H - \Sigma \\
\implies [\mathrm{G}^A]^{-1} = \left[[\mathrm{G}^R]^{-1}\right]^{-1} &= EI - H - \Sigma^\dagger \ , \\
\implies [\mathrm{G}^A]^{-1} - [\mathrm{G}^R]^{-1} &= \Sigma - \Sigma^\dagger = -i\Gamma
\end{aligned}
\tag{A.66}
$$

multiplying with G^R from the left and G^A on the right gives

$$
\mathrm{G}^R - \mathrm{G}^A = -i\mathrm{G}^R\Gamma\mathrm{G}^A \ , \tag{A.67}
$$

which gives us an important NEGF identity for equilibrium. If we now use the matrix A as previously defined we get

$$
A = i[\mathrm{G}^R - \mathrm{G}^A] = \mathrm{G}^R\Gamma\mathrm{G}^A = \mathrm{G}^A\Gamma\mathrm{G}^R \ . \tag{A.68}
$$

This result confirms that A is analogous to the DOS as (from equation A.65) this gives

$$
\mathrm{G}^n = [A]f_0 \ , \tag{A.69}
$$

at equilibrium, as predicted.

If there are interactions of electrons with the surroundings of the channel of that it loses momentum or spin, this can be represented by an additional term, Σ_0 and Σ_0^{in}, such that

$$
\begin{aligned}
\Sigma &= \Sigma_1 + \Sigma_2 + \Sigma_0 \\
\Sigma^{in} &= \Sigma_1^{in} + \Sigma_2^{in} + \Sigma_0^{in} \ .
\end{aligned}
\tag{A.70}
$$

A.2.4 Scattering theory

In the NEGF framework, Σ is given by the boundary conditions. Here, we will look at a simple example to demonstrate the method. We consider a 1D wire (Fig. A.4) featuring discrete points. From tight-binding theory, this wire can be described by a Hamiltonian matrix with energy, ϵ, along the main diagonal and t on upper and lower diagonal, *i.e.*

$$
H = \begin{pmatrix} \epsilon & t & \cdots & 0 \\ t & \epsilon & \ddots & \vdots \\ \vdots & \ddots & \ddots & t \\ 0 & \cdots & t & \epsilon \end{pmatrix} \ . \tag{A.71}
$$

This originates from periodic boundary conditions with a dispersion relation of

$$
E = \epsilon + 2tcos(ka) \ . \tag{A.72}
$$

For this problem we must use open boundary conditions as electrons flow through the system. We consider an incoming wave arrives at one side of the device where it can be reflected or transmitted. This is shown in Figure A.4, where B is the incident wave amplitude, a is the lattice constant for the discrete points, ρ is the reflection coefficient and τ is the transmission coefficient.

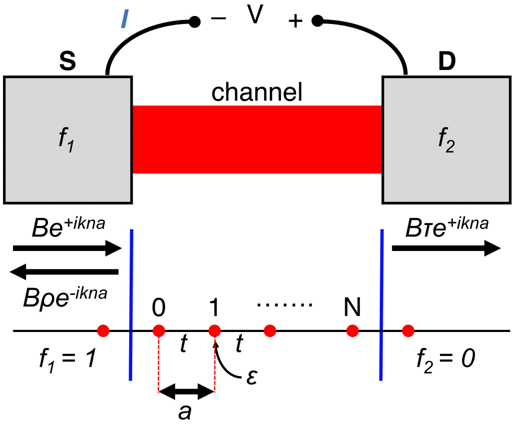


Figure A.4: NEGF formalism for an incident wave on a 1D wire, formed by grid points separated by distance a. Each point has an energy of ϵ and is coupled to its nearest neighbour by t, in accordance with tight-binding theory.

If we go the the end (right side) of the channel, we consider the point inside the boundary and the first point outside, labelled N and $N+1$ respectively. Here, the wavefunction must be

$$\begin{aligned} \psi_N &= B\tau e^{ikNa} \\ \psi_{N+1} &= B\tau e^{ika(N+1)} \\ \implies \frac{\psi_{N+1}}{\psi_N} &= e^{ika} \; . \end{aligned} \tag{A.73}$$

The same procedure on the opposite boundary of the channel (left side), between points -1 and 0 allows us to write

$$\begin{aligned} \psi_0 &= B + B\rho \\ \implies \psi_0 e^{ika} &= Be^{ika} + B\rho e^{ika} \\ \psi_{-1} &= Be^{-ika} + B\rho e^{ika} \\ \implies \psi_{-1} &= \psi_0 e^{ika} + B(e^{-ika} + e^{ika}) \\ \implies \psi_{-1} &= \psi_0 e^{ika} - 2iBsin(ka) \; . \end{aligned} \tag{A.74}$$

Now we must use tight-binding equations:

$$\begin{aligned} E\psi_0 &== t\psi_{-1} + \epsilon\psi_0 + t\psi_{+1} \\ E\psi_N &== t\psi_{N-1} + \epsilon\psi_N + t\psi_{N+1} \; , \end{aligned} \tag{A.75}$$

combining with our previous results (equations A.74) we get

$$\begin{aligned} E\psi_0 &= -2iBtsin(ka) + (\epsilon + te^{ika})\psi_0 + t\psi_{+1} \\ E\psi_N &= t\psi_{N-1} + (\epsilon + te^{ika})\psi_N \ . \end{aligned} \tag{A.76}$$

For a channel consisting of three discrete points ($N = 2$), we can use these results to obtain the following set of matrices when realising that the extra te^{ika} terms picked up at the boundaries correspond to Σ_1 and Σ_2, and the extra term on the left-hand side at point 0 is the inflow (s_1):

$$H = \begin{pmatrix} \epsilon & t & 0 \\ t & \epsilon & t \\ 0 & t & \epsilon \end{pmatrix} \tag{A.77}$$

$$\Sigma_1 = \begin{pmatrix} te^{ika} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \qquad \Sigma_2 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & te^{ika} \end{pmatrix} \tag{A.78}$$

$$s_1 = \begin{pmatrix} -2iBsin(ka) \\ 0 \\ 0 \end{pmatrix} \qquad \textbf{(at } \boldsymbol{N}\textbf{=0)} \ . \tag{A.79}$$

We can now proceed within the NEGF framework to calculate Γ_1

$$\Gamma_1 = i(\Sigma_1 - \Sigma_1^\dagger) = \begin{pmatrix} -2tsin(ka) & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \tag{A.80}$$

$$\Gamma_2 = i(\Sigma_2 - \Sigma_2^\dagger) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -2tsin(ka) \end{pmatrix} \tag{A.81}$$

$-2tsin(ka)$ appears in various places as it is related to the electron velocity in the wire (v), where, from the dispersion relation (equation A. 72) we have

$$\begin{aligned} \hbar v = \frac{dE}{dk} &= -2atsin(ka) \\ \implies -2tsin(ka) &= \frac{\hbar v}{a} \ . \end{aligned} \tag{A.82}$$

We can now rewrite the anti-hermitian outflow matrices as

$$\Gamma_1 = \begin{pmatrix} \frac{\hbar v}{a} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \qquad \Gamma_2 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \frac{\hbar v}{a} \end{pmatrix} \tag{A.83}$$

and we can also rewrite the inflow column vector s_1 as

$$s_1 = \begin{pmatrix} -iB\frac{\hbar v}{a} \\ 0 \\ 0 \end{pmatrix} \tag{A.84}$$

and we can now use $s_1 s_1^\dagger = \Sigma_1^{in}$ (equation A.16) to find

$$\frac{\Sigma_1^{in}}{2\pi} = \begin{pmatrix} \left(\frac{Bhv}{a}\right)^2 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} . \tag{A.85}$$

If we now consider the condition

$$\Sigma_1^{in} = \Gamma_1 f_1(E) , \tag{A.86}$$

as discussed previously, we require

$$B^2 = \underbrace{f_1(E)}_{\text{electrons}} \underbrace{\frac{a}{2\pi hv}}_{\text{1D DOS}} , \tag{A.87}$$

to satisfy this relation. Thus, we can determine the wave amplitude from the open boundary conditions.

A.2.5 Transmission

We now aim to derive an expression for the current within the NEGF framework. We begin by considering inflow and outflow through the first contact which, as previously demonstrated, can be written as

$$I_1 = \frac{q}{h}\mathbf{Tr}\left[\Sigma_1^{in} A \Gamma_1 \mathrm{G}^n\right] . \tag{A.88}$$

Replacing Σ_1^{in} from equation A.86 gives the expression

$$I_1 = \frac{q}{h}\mathbf{Tr}\left[\Gamma_1 (A f_1 - \mathrm{G}^n)\right] , \tag{A.89}$$

which holds as long as contacts can be described by a Fermi function. We can now use the NEGF equations to simplify this expression:

$$\begin{aligned} \mathrm{G}^n = \mathrm{G}^R \Sigma^{in} \mathrm{G}^A &= \mathrm{G}^R(\Gamma_1 f_1 + \Gamma_2 f_2 \mathrm{G}^A) \\ A = \mathrm{G}^R \Gamma \mathrm{G}^A &= \mathrm{G}^R(\Gamma_1 + \Gamma_2)\mathrm{G}^A \\ \implies A f_1 &= \mathrm{G}^R(\Gamma_1 f_1 + \Gamma_2 f_1 \mathrm{G}^A) \\ \implies A f_1 - \mathrm{G}^n &= \mathrm{G}^R \Gamma_2 \mathrm{G}^A (f_1 - f_2) . \end{aligned} \tag{A.90}$$

Substituting this result into the current expression given in equation A.89 results in the following current result

$$I_1 = \frac{q}{h}\mathbf{Tr}[\Gamma_1 \mathrm{G}^R \Gamma_2 \mathrm{G}^A](f_1 - f_2)\ . \tag{A.91}$$

Note that this is the current for a small range of energies, dE. The total current must integrate over all energies such that

$$I_1^{TOTAL} = \frac{q}{h}\int_{-\infty}^{+\infty} \mathbf{Tr}[\Gamma_1 \mathrm{G}^R \Gamma_2 \mathrm{G}^A](f_1 - f_2)\ . \tag{A.92}$$

Comparing this to the well-known Landauer-Büttiker formula [282] for current, we can clearly see that the transmission ($T(E)$) part for the current, *i.e.*

$$T(E) = \mathbf{Tr}[\Gamma_1 \mathrm{G}^R \Gamma_2 \mathrm{G}^A]\ . \tag{A.93}$$

Given that any contact, p, can be described by

$$\Sigma_p^{in} = \Gamma_p f_p(E)\ , \tag{A.94}$$

then the current from that contact (per unit energy) is

$$I_p = \frac{q}{h}\mathbf{Tr}[\Sigma_p^{in} - \Gamma_p \mathrm{G}^n]\ . \tag{A.95}$$

If we now apply this for Af_p and G^n for for any number of contacts, q, we get

$$\begin{aligned} Af_p &= \mathrm{G}^R(\Gamma_1 f_p + \Gamma_2 f_p + \cdots + \Gamma_q f_p)\mathrm{G}^n \\ \mathrm{G}^n &= \mathrm{G}^R(\Gamma_1 f_1 + \Gamma_2 f_2 + \cdots + \Gamma_q f_q)\mathrm{G}^n \\ \implies Af_p - \mathrm{G}^n &= \sum_q \mathrm{G}^R \Gamma_q \mathrm{G}^A (f_p - f_q)\ . \end{aligned} \tag{A.96}$$

This result allows us to write a general expression for multi-terminal devices by returning to equation A.89, which we can now write as

$$I_p = \frac{q}{h}\sum_q \mathbf{Tr}[\Gamma_p \mathrm{G}^R \Gamma_q \mathrm{G}^A](f_p - f_q)\ . \tag{A.97}$$

Relating this to the Landauer-Büttiker formalism gives the trace part of the expression as the transmission between contacts p and q:

$$\overline{T}_{pq}(E) = \mathbf{Tr}[\Gamma_p \mathrm{G}^R \Gamma_q \mathrm{G}^A]\ , \tag{A.98}$$

which can be interpreted as the probability that an electron can transmit from contact p to contact q.

We will demonstrate this point with a simple example: the transmission through a single potential barrier. The problem consists of a 1D wire with a sharp potential barrier of $U_0\delta(x)$ (Fig.

A.5**(a)**]. Here, we have an incident wave, reflected wave and transmitted wave from the potential barrier. Typically, this problem would be solved by considering the continuity of wavefunctions and their derivatives with the Schrödinger equation [211]. Here, we will demonstrate the NEGF method and achieve the same result.

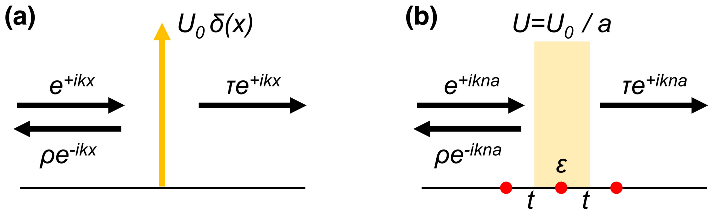


Figure A.5: Schematic representation of transmission through a sharp potential barrier for **(a)** A continuous system and **(b)** A discretised version of the problem.

The problem is first discretised into points of lattice spacing length a (Fig. A.5**(b)**]. Here, the δ function for the barrier must be written as

$$U = \frac{U_0}{a} \,, \tag{A.99}$$

as the lattice produces a barrier of finite width. We next take the discrete lattice and apply the NEGF framework. The barrier contains a single discretized point, giving a Hamiltonian which is just a number;

$$H = \epsilon + U \,. \tag{A.100}$$

Similarly, the matrices Σ_1, Σ_2, Γ_1 and Γ are also single numbers for this problem. More complex examples with n lattice points will have $n \times n$ matrices to solve, however this is easily carried out by a computer. From the previous section, we have

$$\begin{aligned} \Sigma_1 &= te^{ika}, \qquad & \Sigma_2 &= te^{ika}, \\ \Gamma_1 &= \frac{hv}{a}, \qquad & \Gamma_2 &= \frac{hv}{a}, \end{aligned} \tag{A.101}$$

we can now calculate the retarded Green's function for the system:

$$\begin{aligned} \mathrm{G}^R = [E - H - \Sigma]^{-1} &= \frac{1}{E - \epsilon - U - 2te^{ika}} \\ &= \frac{1}{E - \epsilon - U - 2t[cos(ka) - isin(ka)]} \,. \end{aligned} \tag{A.102}$$

Using the dispersion relation given previously in equation A.72, we can write

$$\mathrm{G}^R = \frac{1}{-U - 2itsin(ka)} \,. \tag{A.103}$$

From the derivative of the dispersion relation we can use the relation

$$\frac{\hbar v}{a} = -2tsin(ka) \ , \tag{A.104}$$

which was previously calculated from dE/dk. Thus, the retarded Green's function can be written as

$$\mathrm{G}^R = \frac{1}{-U + \frac{i\hbar v}{a}} \ , \tag{A.105}$$

the transmission can now be computed as

$$\begin{aligned}\overline{T}(E) &= \mathbf{Tr}[\Gamma_1 \mathrm{G}^R \Gamma_2 \mathrm{G}^A] \\ &= \frac{\hbar v}{a} \frac{1}{-U + \frac{i\hbar v}{a}} \frac{\hbar v}{a} \frac{1}{-U + \frac{i\hbar v}{a}} \\ &= \frac{\left(\frac{\hbar v}{a}\right)^2}{U^2 + \left(\frac{\hbar v}{a}\right)^2} \ .\end{aligned} \tag{A.106}$$

If we now consider U over length a (equation A.99), the final transmission expression eliminates all lattice spacings (a):

$$\overline{T}(E) = \frac{(\hbar v)^2}{U_0^2 + (\hbar v)^2} \ , \tag{A.107}$$

which is the standard result which can be obtained with the Schrödinger equation [283].

The simulation technique used in this work is, of course, much more complicated than this example: The matrices are much larger, the computational techniques are optimised for efficiency and scattering mechanisms are included [167]. However, the power of the technique is demonstrated in this simple example.

A.3 Büttiker Probe Scattering

In the NEGF formalism, the carriers cannot redistribute their energy and momentum at scattering events. A technique used to include incoherent scattering and dephasing was first introduced by Markus Büttiker. The basic concept is to include a virtual contact on the device which is analogous to a voltage probe. In the context of the NEGF framework set out previously, additional Σ terms are treated as another contact which can be described by a Fermi function in accordance with equation A.94. Typically these are represented as inscattering and outscattering functions, denoted as Σ_ϕ^{in} and Σ_ϕ^{out} respectively. These can expressed as:

$$[\Sigma_\phi^{in}] = f_\phi[\Gamma_\phi] \qquad \text{and} \qquad [\Sigma_\phi^{out}] = (1 - f_\phi)[\Gamma_\phi] \ , \tag{A.108}$$

where $f_\phi(E)$ is the distribution function of the fictitious probe and Γ_ϕ is the non-coherent rate of electron loss [211]. Typically, the scattering strength for the above functions is modelled by a phenomenological scattering parameter. We can use this framework to express the non-coherent current component flowing from a given contact p:

$$[I_p]_{\text{non-coherent}} = \frac{2e}{h}\int \overline{T}_{p\phi}[f_p - f_\phi]dE \ , \tag{A.109}$$

where

$$\overline{T}_{p\phi} = \mathbf{Tr}[\Sigma_p \mathrm{G}^R \Gamma_\phi \mathrm{G}^A] \ . \tag{A.110}$$

In essence, the virtual contact (Büttiker probe) acts as an external perturbation to the system whereby carriers can be removed, thermalised and reinjected. Consequently, the Büttiker method imitates phonon scattering mechanisms. The model encapsulates important features of dissipative quantum transport. Importantly, the Büttiker-probes can be computed in the same way as real contacts which significantly decreases the computational complexity compared to the self-consistent NEGF method. The full details of this are available in [167]. The technique is found to accurately reproduce the results of the fully consistent NEGF method, including all relevant scattering mechanisms, but is more computationally efficient by many orders of magnitude [166].

Appendix B

InAs/AlSb CB offset

There is a disparity in the calculated InAs/AlSb CB offset between the two simulation methods: nextnano++ and nextnano.MSB. This arises from a difference in the VB offset values given in the simulation databases. The nextnano++ material database quotes the VB offset from Vurgaftman *et al* [197], whereas the nextnano.MSB uses the value from Wei & Zunger [232]. This gives a 2.1 eV CB offset for nextnano++ and 1.85 eV for nextnano.MSB, as shown in Figure B.1**(a)**. It is not clear which VB offset is most accurate, however similar GaSb-based charge storage devices have shown that the localization energies in [197] were systematically overestimated and that the material parameters given by Wei & Zunger [232] more accurately describe the experimental results [284]. The reduction of the InAs/AlSb CB offset will reduce the theoretical storage time of the memory, however it will remain significantly higher than Flash memory as the CB band-offset of 1.85 eV is still extraordinarily large. For example, a barrier energy around 1.9 eV yields a theoretical 300 K storage time in the region of $10^9 - 10^{14}$ years [196].

We next consider the effects this may have on P/E cycling. As all of the resonant-tunnelling physics occurs within 1 eV of the InAs CB minima, it is unlikely that the small change in offset will have a detrimental impact on the resonant tunnelling mechanism. However, the change in barrier height shifts the position of the ground states of the QWs within the TBRT region (QW_1 and QW_2). In order to demonstrate this, simulations were conducted on a similar (TBRT) structure using the two different material databases, after which the CB minima of InAs is used to normalise the results [Fig. B.1**(a)**]. The difference between the QW ground states for the two methods (cyan-line with probability for nextnano.MSB and red-dot dash line for nextnano++) is less than 5 meV [Fig. B.1**(b)**]. This extremely small shift should not change the resonant tunnelling current-density results in any significant way. The most likely difference is a < 5 mV shift in the resonant peaks of the current-density plot. Consequently, one can assume that the disparity will not have any adverse effect on memory operation regardless of which VB offset is the correct one.

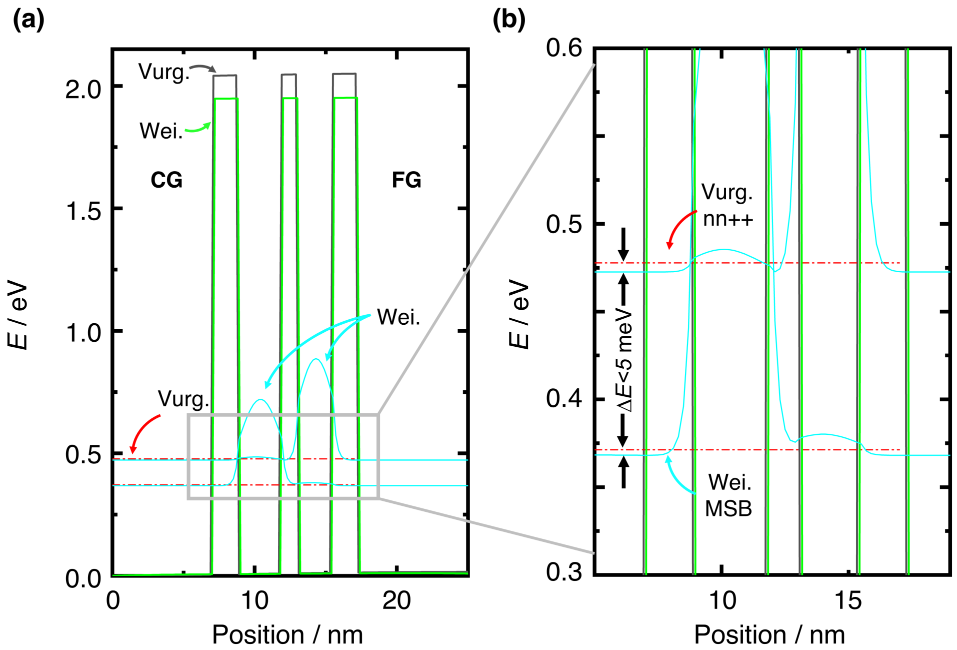


Figure B.1: Comparison of nextnano++ (grey line CB) and nextnano.MSB (green line CB) simulations for the TBRT region which use VB offsets from Vurgaftman and Wei & Zunger respectively. The cyan line represents the energy levels for the QWs of Wei & Zunger with $|\psi|^2$ included to indicate which QW the energy corresponds to. The red dot-dashed line in the QW energy levels for the Vurgaftman CB offset. **(b)** is a magnified representation of **(a)** to demonstrate the difference in energy states for the two methods.

Appendix C

UVL Mask Layout

C.1 Single Devices

C.1.1 Mesa definition

This masking sector has chrome features on a clear glass background [Fig. C.1**(a)**]. For single devices, the widths range from 40 μm to 120 μm and determine the mesa areas to be protected during mesa isolation. This process is explained in detail in subsection 5.4.1.

C.1.2 Source-drain fabrication

The following masking layer is used to define source and drain areas of this sample. As depicted in Figure C.1**(b)**, this is carried out using glass window features on either end of the mesa. These windows are slightly wider than the mesa areas for alignment purposes. This lithographic pattern is used to etch down to gain access to the device's channel and thus define the control gate of the memory.

C.1.3 Source-drain contact fabrication

The source-drain contact layer of the photomask [Fig. C.1**(c)**] features open glass windows on a chrome background which are placed within the mesa and the source-drain fabrication windows. Thus, this allows metal source-drain contacts to be deposited prior to gate dielectric deposition. The backgate sector is also open in this mask, allowing for backgate metallisation within this process stage.

C.1.4 Control gate/wordline contacts

This masking sector [Fig. C.1**(d)**] consists of a single open glass window on each device which overlaps with the previous source-drain contact layer. It's purpose is to allow for metallisation of the control gate (and entire wordline on the arrays) which overlaps the source-drain contacts when separated by the gate dielectric. This is critical for memory performance in NORMALLY-OFF designs, as detailed in subsection 4.8.

C.1.5 Oxide etching windows

Open windows (glass) with a chrome background provide openings within the metal control gate, source and drain terminals [Fig. C.1**(e)**]. This allows for the selective chemical removal of insulating material to reveal the contacts in these windows.

C.1.6 Lifting layer

Contact lifting layers were formed using this masking sector as described in subsection 5.4.9. 600 $\mu m \times 100\ \mu m$ chrome rectangles preserve the underlying resist during the positive lithography. The resist is positioned to have overlap with one edge of the device [Fig. C.1**(f)**] for the purposes described in 5.4.9.

C.1.7 Final contact layer

This masking layer provides glass openings for the deposition of final contacts on the device, including windows of at least 200 $\mu m \times 160\ \mu m$ windows for wire bonding and contact probing. The layer creates a path from the bonding pads to the CG, S and D contacts previously made accessible through oxide etching [Fig. C.1**(g)**].

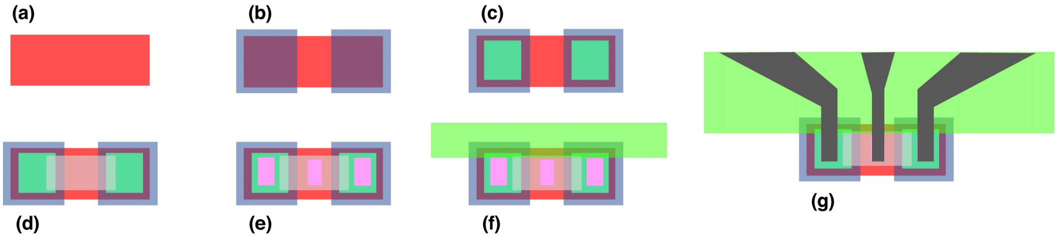


Figure C.1: Detailed schematic of different mask sectors of *Novel Nibble*, each representing a UVL process step relating to single memory device fabrication. **(a)** Mesa definition. **(b)** Source-drain fabrication. **(c)** Source-drain contact fabrication. **(d)** Control gate/wordline contacts. **(e)** Oxide etching windows. **(f)** Lifting layer. **(g)** Final contact layer.

C.2 Backgate

C.2.1 Mesa definition

Chrome features preserve the backgate region during the mesa definition of the devices and arrays [Fig. C.2**(a)**].

C.2.2 Source-drain fabrication

An open glass window allows the backgate region to be etched during the source-drain fabrication [Fig. C.2**(b)**]. Crucially, this ensures that the backgate access on the array and backgate are the same starting level during the next step.

C.2.3 Back gate fabrication

A similar open glass window is used to allow for etching to the BG region [Fig. C.2**(c)**] as detailed later in subsection 5.4.3.

C.2.4 Source-drain contact fabrication

Contacts are added at this stage using a similar open window [Fig. C.2**(d)**] to minimise the backgate material (InAs) to atmospheric conditions before metal is deposited. This reduces the effect of oxidation on the backgate surface and therefore improves contact resistance at the InAs/metal heterojunction.

C.2.5 Oxide etching windows

Again, a clear window is used to access to backgate at this stage [Fig. C.2**(e)**]. This is used to chemically remove oxide from the surface in order to reveal the BG terminal for later processing.

C.2.6 Final contact layer

This layer is added to thicken the initial contact layer with a second metal deposition in order to improve chances of wire bonding success [Fig. C.2**(f)**].

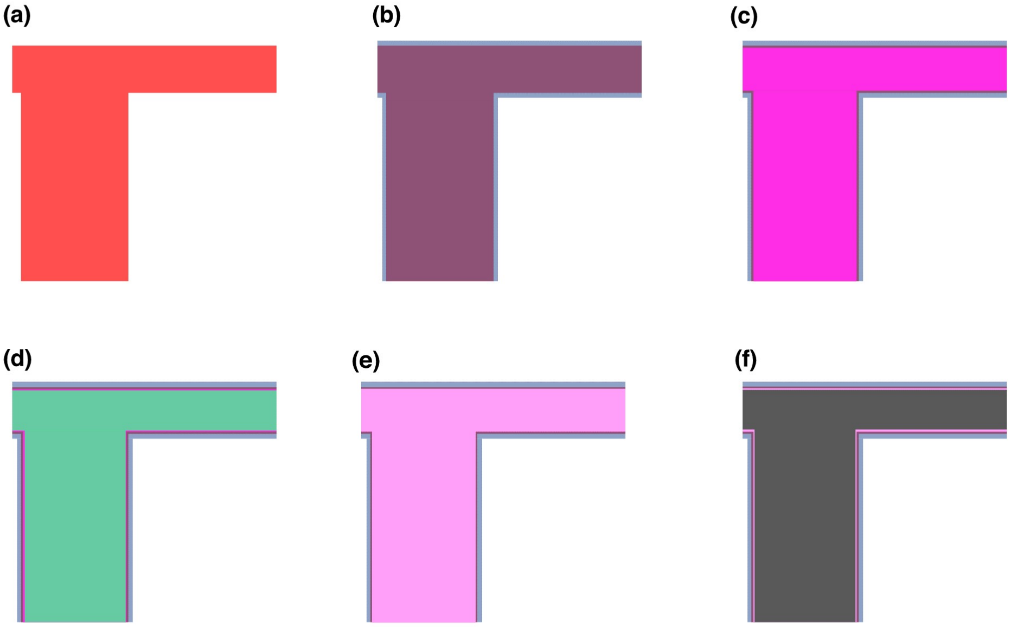


Figure C.2: Detailed schematic of different mask sectors of *Novel Nibble*, each representing a UVL process step relating to top-access backgate fabrication. **(a)** Mesa definition, **(b)** Source-drain fabrication. **(c)** Backgate fabrication. **(d)** Source-drain contact fabrication. **(e)** Oxide etching windows. **(f)** Final contact layer.

Appendix D

Processing details

D.1 UVL

For a reliable resist coating, the sample should first be at room temperature before resist application. Secondly, ample photoresist adhesion between the sample substrate and the resist is crucial, particularly for multiple lithography processes with many alignments. High quality resist sample adhesion results in minimised dark erosion and undercut, enhanced edge quality and improved feature size. The main factors in resist adhesion are the substrate material itself (hydrophilic or hydrophobic, Fig. D.1) [285], and the formation of water on the sample surface due to atmospheric humidity [286]. For the resists used in this work, the optimum relative humidity is 30 to 50% (optimum 43%), with coatings becoming increasingly difficult and near-impossible around 70% [287]. For the most part, the air moisture was around 40% during memory fabrication.

In this work, resist is applied on the surface of the same sample many times. During the process, the material occupying the uppermost layer changes depending on the processing step and location on the surface (x-y). The surface materials are: In(Ga)As, Al_2O_3 (or HfO_2) and SiO_2. III-V materials are naturally hydrophilic, therefore the soft-baked resist adheres to the surface [288]. The same can be said for the Al_2O_3 and HfO_2 dielectric layers. SiO_2, on the other hand, is sometimes hydrophobic and can produce resist adhesion issues. For the most part, SiO_2 is hydrophilic due to the presence of the silanol (Si-OH) groups on the surface. However, the silanol (Si-OH) groups can often chemically react with various reagents to render the silica hydrophobic [289]. For SiO_2 deposited using the PECVD technique (as used in this work), the silanol groups should be present on the surface when exposed to atmospheric conditions. Assuming that the Si-OH groups are not removed through heating or have reacted with contaminants before resist application, the hydrophilic surface will reliably accommodate the resist coating. This was systematically tested during process calibration. It was found that S1813 and LOR-3A resist coatings are of high quality on flat, featureless SiO_2 (on Si, PECVD) test samples. Fortunately, this means that the use of SiO_2 in this work does not require any additional layers to improve resist adhesion.

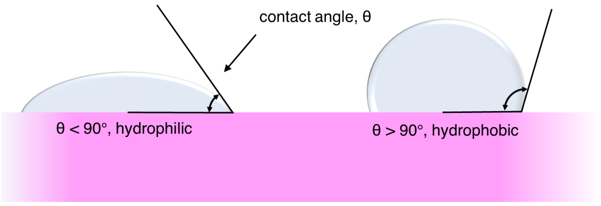


Figure D.1: Comparison of hydrophilic and hydrophobic surfaces including contact angle.

D.1.1 S1813

Table D.1 lists the spinning parameters for the Microposit S1813 positive photoresist. This resist is used for patterning the memory structures in preparation for etching processes and the contact lifting layer. The resist is dispensed onto the sample using a pipette. After the spin coating is complete, the coated sample is placed face-up on a acetone smeared cleanroom wipe for 20 s. This step allows acetone to penetrate into the edges of the sample to reduce edge bead size whilst also cleaning any unwanted resist from the backside of the sample. Any remaining resist on the backside is then carefully wiped away using an acetone soaked cotton bud. For single layer S1813 coatings, we perform a 120 s soft-bake on a 115 °C hotplate followed by a 3 minute cool-down period. The resulting resist film thickness was measured to be around 1250 nm for this process with less than 50 nm of thickness variation across the sample.

UV exposure for the masking patterns was performed in soft-contact mode. Development consisted of gentle stirring in room-temperature MF-CD-26[1] for 60 s followed by a stir in de-ionised water for 60 s and N_2 blow drying.

Later resist stripping following etching processes consisted of a 3 minute immersion in room-temperature acetone which is then agitated in an ultrasonic bath at 50% power. This is followed by a 2 minute rinse in IPA terminated with N_2 blow drying. If metal contacts are exposed on the sample surface, the ultrasonic bath is substituted for gentle stirring in order to prevent contact delamination from the vibrations on the chip. If this process is found to leave resist residue on the surface, O_2 plasma ashing (subsection 3.3.2) is used to clean the sample.

Heating the resist above certain threshold temperatures cross links the polymers and makes it more stable, less removable and strengthens the resist against a range of chemicals. For the wet chemical etch described in Appendix D.2, Microposit MF-319 is used as an etchant. As this is also a TMAH-based developer, which is now being used in an acid wet bench, UV light is incident on the sample meaning a soft-baked resist would gradually develop across the entire sample causing the resist to perish in the TMAH solution. However, a 5 minute post-bake at 125 °C causes the resist to become resistant to TMAH-based solutions, even when the entire sample is exposed

[1] MF-CD-26 is a solution containing 2.4% tetramethylammonium hydroxide (TMAH) and water.

Table D.1: Spin parameters for photoresist spin coatings.

	Step	Duration /s	Acceleration /rpms^{-1}	Speed /rpm
S1813	1	5	500	1500
	2	60	1500	6000
	3	5	1500	500
LOR-3A	1	5	500	1000
	2	30	1000	3000
	3	5	500	1000

to UV light afterwards. Fortunately, this bake temperature does not render the layer resistant to acetone, and therefore the resist can be stripped post-etch as previously described.

At even higher temperatures, the cross linking of resist polymers causes the material to become very chemically resistant. At hard-bake temperatures above 200 °C, S1813 is unharmed by the most aggressive stripping chemicals, including acetone and warmed REMOVER 1165[2]. Later, this will be used to produce contact lifting layers on the sample.

D.1.2 LOR-3A/S1813 bilayer for metallisation

S1813 is successfully used to pattern the sample for etching and lifting layer processes. However, this resist alone is not ideal for patterning metallic features due to the outward incline towards the exposed/developed areas of the resist. The outward incline allows the deposition material to settle on the sidewalls of the resist. Consequently, the resist is sealed from the removing chemicals (acetone) which leads to a problematic liftoff. To resolve this issue and produce metal features of high-resolution, a lift-off layer such as LOR-3A is used when metal is required.

For metallisation processes, LOR-3A is first spun on the sample surface (parameters are presented in Table D.1) and the backside is carefully cleaned with Remover 1165. The lift-off resist in then softbaked on a hotplate at 170 °C for 5 minutes and is then allowed to cool to room temperature. A layer of S1813 (as described previously) is then spun atop of the LOR-3A layer before a standard soft-bake (120 s, 115 °C) is performed. Once the sample is at room temperature, it is exposed in soft-contact mode for 2.0 s using the desired masking pattern. Next, a 45 s development in MF-CD-26 with gentle stirring is carried out before rinsing in de-ionised water and an N_2 blow dry. The sample is then post-baked at 125 °C for 5 minutes. The purpose of this bake is to produce a more developer-resistant S1813 layer, without changing the properties of the LOR-3A layer residing underneath. Finally, the resist layers are developed for another 60 s (then de-ionised water rinsing and drying). The difference in development rates due to the post-bake leaves an overlap of the S1813 layer which prevents sidewall deposition - producing a very

[2] MICROPOSIT REMOVER 1165 is a mixture of pure organic solvents specifically formulated to remove all Shipley photoresists. It is particularly recommended for use in applications where the photoresist has seen high temperatures, strong etchants, or other harsh processing conditions.

reliable and high-resolution lift-off process.

D.2 Source-drain chemical etching

InAs etching is carried out using a solution of citric acid, hydrogen peroxide and de-ionised water. In this work, all citric acid etch processes are carried out at room temperature using a 1:3:1 volumetric ratio of citric acid, hydrogen peroxide and de-ionised water respectively. The etch rate of InAs in this solution is predicted to be around 50 nm/min [290]. The chemical process of the etching mechanism has been shown to proceed by an oxidation-reduction reaction at the material surface by the H_2O_2, and subsequently the oxidised material is dissolved by the acid [290]. For Sb-based alloys such as GaSb and AlSb, the etch rate is significantly slower. The etch rate of InAs over GaSb has a selectivity of at least 100, whilst an $Al_0.5Ga_0.5Sb$ alloy is almost impervious to all citric-acid solutions (< 1 Å/min) [290], so we can assume AlSb will provide a good etch-stop layer to the InAs etch.

Microposit MF-319, much like MF-CD-26, is a TMAH-based solution typically used for photoresist development. Both solutions contain a small percentage (2-3 %) of TMAH with water. MF-319 became the etchant of choice in this work for consistency with the literature on AlSb/GaSb etching selectively over InAs [291]. In the case of etching of AlSb, a surface layer rich in Sb is built, whereas no elemental Sb appears when GaSb is etched. This difference in the etching behaviour can be explained by the difference in reaction heat between Al and Ga which is also responsible for the fact that AlSb disintegrates rapidly under ambient atmosphere whereas GaSb is stable. The Al immediately oxidizes on contact with OH- ions, whereas Sb reacts much more slowly so there are atoms left over without reaction partners. As such, an Sb layer is built on the AlSb surface. When GaSb is etched, the reactivities of the III and the V elements are more balanced. Ga is not as reactive as Al and most Sb can react to produce oxide formations which are then dissolved in the etchant.

The total time required for complete removal of the AlSb layers is a non-linear function of layer thickness and cannot be calculated straightforwardly from etch data. It has been found that the time to remove a layer of AlSb entirely (t) consists of the time t_{ox} to oxidize all Al from AlSb, the time t_{etch} to dissolve the aluminium oxide formed during oxidation with additional time t_{Sb} to remove any remaining Sb deposited on the underlying InAs layer, *i.e.*

$$t = t_{\text{ox}} + t_{\text{etch}} + t_{\text{Sb}} = t_{\text{ox}} + \frac{a_{\text{AlSb}}}{r} + t_{\text{Sb}} \tag{D.1}$$

where a_{AlSb} is the AlSb layer thickness and r = 0.7 nm / s is the steady post-oxidation AlSb etch rate [291]. In this work, only a few monolayers of AlSb for the each barrier of the TBRT is removed in each step. It is known that these layers will oxidise almost instantaneously upon being exposed from the former citric etch step [192] (*i.e.* t_{ox} is finished before the etch even begins). It is observed

Table D.2: Etch steps for selective layer-by-layer S-D etching of ULTRA**RAM**™ memories. Each step is followed by a 60 s rinse in de-ionised water. Etch thicknesses for each step can be deduced from Figure D.3.

Step	Etchant	Duration /s	Layers removed
1	20	citric	InAs FG
2	10	MF-319	AlSb barrier 3
3	20	citric	InAs QW_2
4	10	MF-319	AlSb barrier 2
5	20	citric	InAs QW_1
6	10	MF-319	AlSb barrier 1

that the AlSb barriers dissolve in a just a few seconds in the TMAH solution.

The exact parameters of the etch are provided in Table D.2, whereby the etching steps (1-6) are carried out in order and each is followed by a 60 s rinse in de-ionised water. These etching times were carefully calibrated prior to memory processing, the results of which are presented below (Fig. D.2).

Although the selectivity for the InAs and AlSb etches is extremely high, it was found that the etchant dip times require precise calibration. A test-dip was carried out on a material growth of the ULTRA**RAM**™ structure (*v2.1*, GaAs substrate growth). Based on cross-sectional TEM measurements, the channel layer is buried around 30 nm from the wafer surface. Etches with alternating 20 s dips in the citric/MF-319 solutions similar to those presented in Table D.2 were carried out. Step height AFM analysis of the etch [Fig. 5.11 **(d)-(e)**] measured the etch depth as 32 nm. Considering the degree of error in the TEM and AFM measurements, we have confidence that this etch has terminated within the channel layer as intended.

Most of the etched surface is smooth, however there are small holes of around 15 nm depth on the channel surface (Fig. D.2**(e)**, left side). It is likely that these features are the cause of the cloudy appearance of the etch when imaged with an optical microscope [Fig. D.4**(b)**]. To investigate this further, the etch was repeated where the individual dip times were varied across three separate samples [Fig. D.2 **(a)-(c)**]. The 10 s and 20 s dip times produced the same etch depth (within uncertainty). This was expected and indicates that there is material selectivity of the etchants used, whereby each etch step in halted by the underlying layer. Unexpectedly, the 30 s dip time produces a much deeper etch in which step height analysis places us near to the InAs BG layer (*v2.1* GaAs growth used in this test). Moreover, the etch roughness is significantly increased and can be easily observed from an optical microscope [Fig. D.2**(c)**]. Based on these results, it is concluded that the holes appearing on the etch surface allow liquid etchant to penetrate into the layer below which is a non-selective material (GaSb). As we increase the time exposed to the etchant, the holes become larger, resulting in a rougher-looking surface without changing the etch depth. Eventually, the widening of the etchant openings causes disastrous lift-off of the layer, in which the channel is removed due to the elimination of its underlying material despite being resistant to the etching chemicals.

The most probable explanation of the emergence of holes in the etched surface is that they coincide with defects and/or threading dislocations in the crystal, which are known to cause etch pits [292]. Consequently, with a very thin channel layer, it is plausible that the etch pits are deep enough to extend into the underlying material layers. Indeed, the 15 nm AFM-measured pit depth is thicker than the 10 nm n-InAs layer of the channel used in Samples A, B and C.

Sample A (v2.1, n-InAs channel memory), with material layers grown on GaAs substrate used this etching procedure. However the channel conductivity demonstrated some instability during memory characterisation. The channel etch quality was greatly improved by substituting the TMAH etchant for one which is less reactive with the GaSb but nevertheless etches AlSb effectively. Buffered oxide etchant (BOE) was identified as a suitable candidate for a direct substitution (*i.e.* the dip times are used as in Table D.2). This etchant substitution was used for all later processes and resulted in stable, Ohmic channel contacts for device testing.

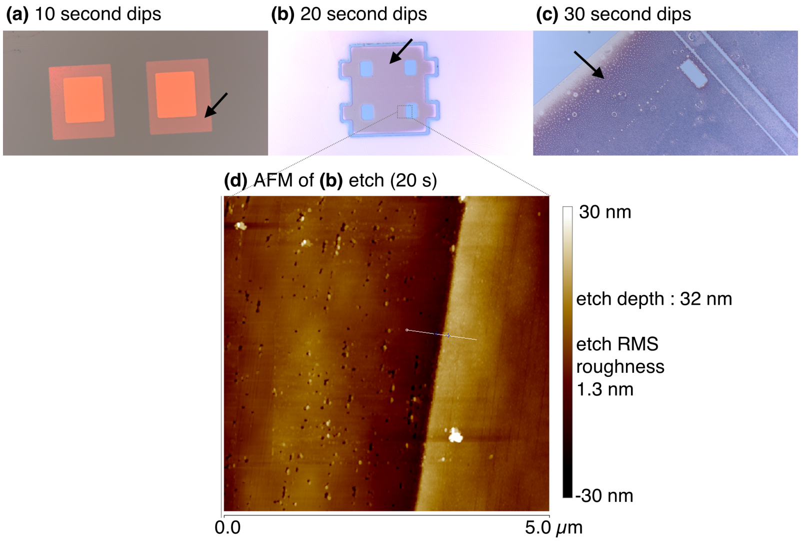


Figure D.2: Results of S-D etching calibrations. **(a)-(c)** Optical microscope images of the etch, where arrows indicates the etched layer. **(d)** AFM of **(b)** etch where the left hand side is the channel (etched) layer and the opposite side is the wafer surface. Etch roughness measurements exclude holes (etch pits).

D.3 Ti-Au sputtering process and post-metallisation resist liftoff

For successful lift-off of LOR-3A/S1813 bilayer resist after metallisation via sputtering or thermal evaporation, the edges of the sample are carefully scraped with a scalpel to remove any metal that is sealing the edges. This allows the remover to attack the LOR-3A lift-off resist from the outer sample edge. It is then placed in 75° C R-1165 for 10 minutes followed by a gentle pipette splashing of R-1165. Reliable lift-off under these conditions requires a deposited film thickness not exceeding 60 nm and 200 nm for sputtering and thermal evaporation respectively.

D.4 Oxide etching process calibration

Resist windows were formed on the surface before etching at various BOE dip times. Confirmation of material removal was carried out using EDX measurements to map the change in Al and O species across the sample. As shown in Figure D.3, a 60 s dip time was sufficient to remove all oxide material and this etch produced high-quality features (no visible underetch). Therefore, all oxide chemical etches in this work are carried out by a 60 s BOE dip followed by a 60 s rinse in de-ionised water.

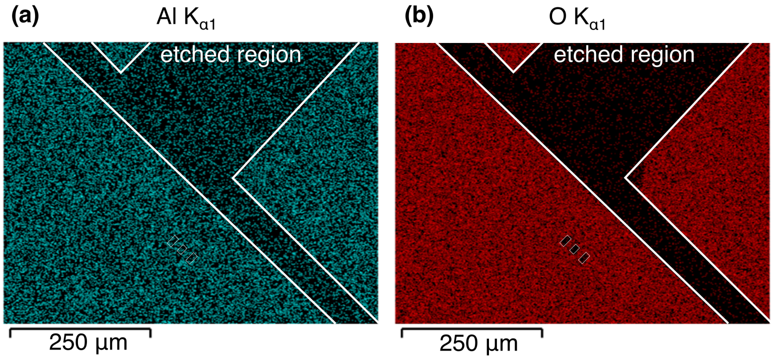


Figure D.3: EDX species mapping on the sample surface for BOE etching of SiO_2 (PECVD) and Al_2O_3 (ALD) to confirm the etching of material within the patterned features for **(a)** Al and **(b)** O elements.

For the ULTRA**RAM**™ process, the underlying material for the etch is a noble metal (gold). Thus, exposed contacts are unharmed during the etch. However, if a TiO_2 or Al_2O_3 adhesion promoter is not used, the etchant can seep into the SiO_2/Au interface in the CG/WL regions. This causes the etchant to damage the entire WL/CG region, as it is not easily rinsed away once seeping between layers. From here, the etchant slowly propagates into the underlying Ti adhesion layer causing irreparable damage to the CG/WL contacts and, in some cases, lifting them off entirely.

Appendix E

Threshold window analysis

The simulation of the program cycle outlined in subsection 4.6.2 assumes that the limit of charge storage on the device FG comes from the capacitance. However, this is a classical interpretation based on the assumption that there are many available electron states in the FG. In general, the FG storage layer of a conventional FG memory (*i.e.* flash) is fabricated from polysilicon or metal [271], so the classical interpretation holds. However, for the thin InAs FG layer used in ULTRA**RAM**™ devices presented in this thesis, this may not be the case. To investigate this, we will calculate the maximum amount of charge that can be stored on the FG under the new assumption that the DOS of the FG is the limiting factor of the program cycle and determine what the threshold window shift (ΔV_T) should be under this condition.

We begin with the Poisson-Schrödinger simulation results of the ULTRA**RAM**™ structure in the TBRT and FG region, shown in Fig. E.1. The 10 nm InAs FG layer forms a QW with discrete energy levels, two of which are within the energy range of the ground state tunnelling energies for the TBRT region. The energy-shifted WF probability, $|\psi|^2$, for ground state and first excited state of the FG are depicted by solid red and green lines respectively. The dotted lines represents the energy alone, which are labelled $E_{FG}^{(0)}$ and $E_{FG}^{(1)}$ for the ground and first excited states in the FG respectively. Likewise, the ground state energies in the TBRT for QW_1 and QW_2 are presented in a similar fashion with blue and magenta lines respectively, and are labelled $E_{QW1}^{(0)}$ and $E_{QW2}^{(0)}$ accordingly.

It is known from detailed NEGF calculations (Section 7.2) that the transmission probability through the TBRT has its first peak when the energy reaches the ground state of QW_1[1]. Therefore, it is unlikely that electrons with energies exceeding the resonant tunnelling energy of $E_{QW1}^{(0)}$ will contribute to the non-volatile state of the programmed logic. In the FG, the first excited state exceeds that of the QW_1 ground state, such that electrons occupying the FG must be in the ground state, *i.e.*

$$E_{QW1}^{(0)} < E_{FG}^{(1)} , \tag{E.1}$$

[1] $E_{QW1}^{(0)}$.

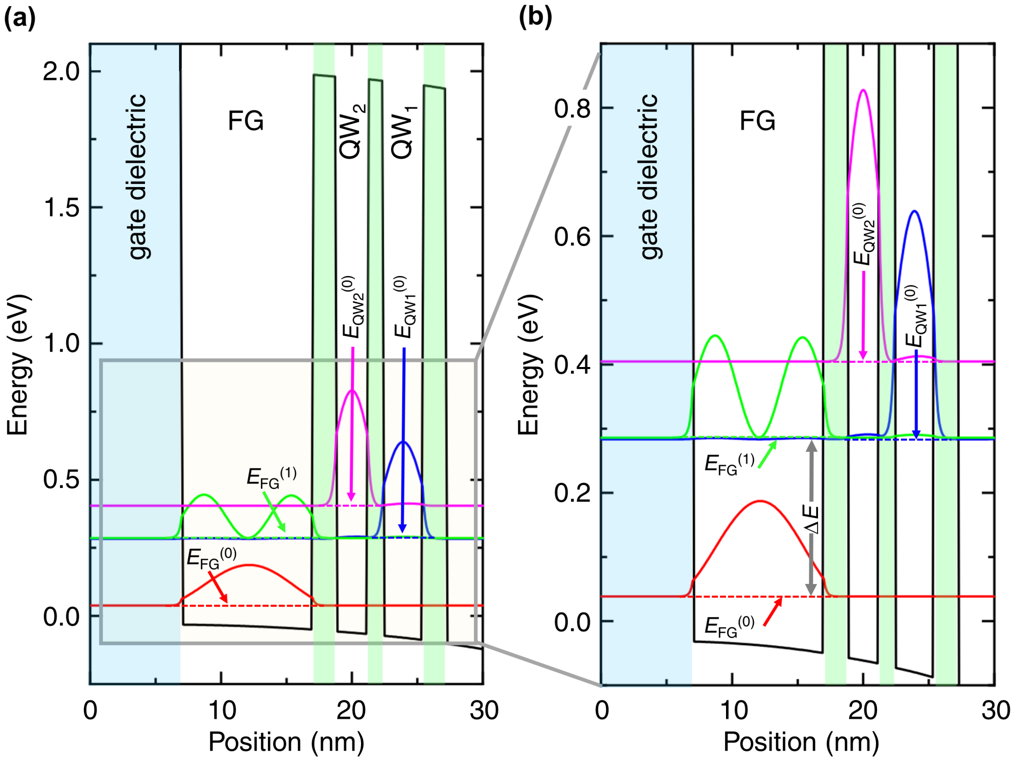


Figure E.1: **(a)** Poisson-Schrödinger simulation results of the ULTRA**RAM**™ structure in the TBRT and FG region. Solid lines are energy-shifted WF probabilities, $|\psi|^2$, for the ground state and first excited state in the FG (red and green respectively). In the TBRT region, the solid blue and magenta lines are energy-shifted $|\psi|^2$ for the ground states of QW_1 and QW_2 respectively. Dashed lines indicate the energy level beneath the non-zero WF probability. **(b)** A magnified version in the energy range around the tunnelling energies.

which can be seen clearly in Fig. E.1**(b)**. Consequently, all electrons in the non-volatile P state must be within the energy range of the two QW ground states for the FG and QW_1. This is labelled ΔE in Fig. E.1**(b)** where

$$\Delta E = E_{QW1}^{(0)} - E_{FG}^{(0)} \, . \tag{E.2}$$

With this in mind, we can proceed to calculate the total charge occupancy of the programmed FG starting with the 2D DOS previously outlined in Section 4.1:

$$g_{2D}(E) = \frac{m_e^*}{\pi\hbar} \, , \tag{E.3}$$

where m_e^* =0.023m_e, the effective electron mass in InAs [197]. For the FG storage this gives a total number of states, n_{2D}, as

$$n_{2D}(E) = \int_{E_{FG}^{(0)}}^{E_{QW1}^{(0)}} g_{2D}(E) = \int_{E_{FG}^{(0)}}^{E_{QW1}^{(0)}} \frac{m_e^*}{\pi\hbar} = \frac{m_e^*}{\pi\hbar}\left(E_{QW1}^{(0)} - E_{FG}^{(0)}\right) = \frac{m_e^*}{\pi\hbar}\Delta E \, . \tag{E.4}$$

Table E.1: Comparison of 2D DOS calculated maximum ΔV_T with measured non-volatile ΔV_T for samples C and D (10 nm InAs FG).

	Sample C	**Sample D**
Gate dielectric	Al_2O_3	HfO_2
Thickness / nm	15	15
k value	$\sim$7	$\sim$16
Calculated max. ΔV_T / V	0.37	0.14
Measured ΔV_T / V	0.32	0.13

The maximum electron density, n_e, can then be calculated using the energy Eigenvalues from the Poisson-Schrödinger calculations of Fig. E.1 with the above expression (degeneracy factor of two included):

$$n_e = \frac{2m_e^*}{\pi\hbar}\Delta E = 4.7 \times 10^{12}\ \text{cm}^{-2}\ . \tag{E.5}$$

This corresponds to a maximum FG charge, Q_{FG}, of

$$Q_{FG} = e \times n_e = 7.5 \times 10^{-8}\ \text{C/cm}^2\ , \tag{E.6}$$

where e is the charge of an electron. This value for maximum Q_{FG}, based on the 2D DOS of the QW FG, can be used to calculate the maximum non-volatile threshold voltage shift of an ULTRA**RAM**™ memory cell from the equation previously given in 2.2.1 (equation 2.3):

$$\Delta V_T = \frac{Q_{FG}}{C_{FG}}\ , \tag{E.7}$$

where C_{FG} is the capacitance between the FG and the CG, *i.e.* the capacitance of the ALD gate dielectric. Using the C_{FG} of the gate dielectrics of Al_2O_3 and HfO_2 films for samples C and D respectively, we can predict the maximum ΔV_T values. The results are presented in Table E.1 alongside the non-volatile ΔV_T measurement from Chapter 6. They offer some explanation of the unexpected decline in the non-volatile window when implementing a higher-k gate dielectric in sample D. Indeed, it is clear that the DOS in the FG should be improved in order to enhance the ΔV_T window, either by changing to a material with higher effective mass or thickening the InAs layer. Alternatively, the dielectric layer capacitance could be reduced to extend the memory window in accordance with equation E.7.