Lancaster University

Scalable Fabrication, Characterisation and Critical Interface Simulation of ULTRA**RAM**™ Memory Devices

Xiuxin Xia

Supervisor:

Prof. Manus Hayne

Department of Physics

A thesis submitted for the degree of

Doctor of Philosophy

July, 2025

Scalable Fabrication, Characterisation and Critical Interface Simulation of ULTRA$\mathbf{RAM}^{\text{TM}}$ **Memory Devices**

Xiuxin Xia July, 2025

Abstract

In this work, the design, fabrication and characterisation of scalable ULTRA$\mathbf{RAM}^{\text{TM}}$ aiming at 50 nm gate dimension are reported, in conjunction with simulations of the interface alloying of the triple-barrier resonant-tunnelling (TBRT), aiming at the scaling-down of the feature size of the memory devices for commercialisation.

ULTRA$\mathbf{RAM}^{\text{TM}}$ is a non-volatile, III-V based compound-semiconductor, charge-storage memory in which the oxide barrier that separates the floating gate from the channel in flash is replaced by the TBRT structure that exploits the 2 eV band offset between InAs and AlSb. This empowers ULTRA$\mathbf{RAM}^{\text{TM}}$ to operate at low voltage and high speed, with high endurance, low disturbance and ultra-low switching energy.

Multiple batches of memory devices were fabricated alongside the optimisation of the self-aligned design. Two key improvements shape the final processing flow. The introduction of all-dry etching enables the nanometre scalability of the memory design for the first time and the employment of laser interferometry in end-point detection allows enhanced accuracy of etching to the 10 nm-thick channel. The characterisation of the as-fabricated devices demonstrated stable performance with memory windows of ~ 10 μA at room temperature and a minimum pulse duration 5 ms was obtained, indicating the success of improvements made in the refining of the fabrication flow. A lingering issue with the gate-drain leakage through dielectric due to the problematic gate metal that arose during the legacy method is solved, and a synchrotron X-ray nano-probe analysis was conducted in an attempt to understand the failure mechanism of the devices.

A resistive-gap hypothesis is proposed to explain the poor performance observed

in as-fabricated devices. A compact device design is then proposed to address the issue and preliminary validation of the fabrication flow shows promising results where the resistive gap is substantially reduced from $\sim$ 5 μm to 385 nm, a factor of ten smaller than the previous design. This is expected to further improve the memory readout performance.

Simulation by nextnano software using the multi-scattering Büttiker (MSB) probe formulism was carried out to investigate robustness of TBRT operation against intermixing after concerns were raised as a result of transmission electron microscopy (TEM) images. All scenarios, including alloying at both barrier and quantum well layers, show only slight impact on the memory operation voltage but no undermining of the TBRT function, evidencing the resilience of the TBRT structure against the growth imperfections or fluctuations, in strong support for the ULTRA**RAM**™ concept. The interplay of these findings indicates ULTRA**RAM**™'s potential as a viable emerging universal memory and its way to practical application in the not too distant future.

Acknowledgements

I would like to express my sincere gratitude to Professor Manus Hayne for his invaluable support, guidance, and patience throughout the course of my PhD studies. His encouragement, particularly during the challenging phases of scaling design, as well as his insightful discussions, have been instrumental to this work.

My heartfelt thanks also go to Dr Peter Hodgson for his continuous efforts in improving wafer growth and optimisation, a fundamental aspect of this study. I am especially grateful to Dr Dominic Lane, Dr Serdar Tekin, Dr Samuel Jones, Dr Jonathan Hall, and Gizem Acar for the many hours spent in the cleanroom and for their invaluable discussions and technical assistance. I extend my sincere appreciation to Garry Vernon for the timely support provided with cleanroom equipment.

I am grateful to Professor Benito Alén for granting me the opportunity to be part of the QUANTIMONY project. I would also like to thank Professor Richard Beanland and Dr Francisco Alvarado for their support with TEM analysis and for warmly welcoming me during my secondment at the University of Warwick. In addition, I appreciate the valuable assistance of Dr Stefan Birner and Takuma Sato with nextnano software support, which was essential to the development of the simulation work.

Special thanks to Dr Samuel Jarvis for the collaboration on XPS measurements, and to Dr Richard Wilbraham for FIB-SEM imaging support. I am also grateful to all collaborators at ESRF for facilitating the beamline measurements. My thanks extend to the staff and fellow students in the Department of Physics, as well as across the project, for their collaboration and support throughout the research.

Finally, I am deeply thankful to my family for their unwavering support, understanding, and patience over the past four years. To them, I am profoundly grateful.

Declaration

I declare that, except where specific reference is made to the work of others, the work presented in this thesis is, to the best of my knowledge and belief, original and my own work. The material has not been submitted, either in whole or in part, for a degree at this, or any other university.

Xiuxin Xia, July 2025

Publications and Presentations

Articles

Critical Interface Simulation of ULTRA**RAM**™ Memory Devices, in preparation.

Patent-pending

M. Hayne, P. Hodgson and X. Xia, Improved Memory Device, P00579, submitted. (2024)

Conferences

- Oral presentation at UK Semiconductors, July 2022, Sheffield, UK, Scaling of ULTRA**RAM**™ III- Sb charge-storage devices for non-volatile random access memories.
- Oral presentation at WOCSDICE-EXMATEC, May 2023, Palermo, Italy, An optimised fabrication flow for scaling of ULTRA**RAM**™ devices.
- Oral presentation at UK Semiconductors, July 2023, Sheffield, UK, Sub-micron scaling of ULTRA**RAM**™ III- Sb charge-storage devices for non-volatile random-access memories.
- Oral presentation at UK Semiconductors, July 2024, Sheffield, UK, ULTRA**RAM**™: Advances in scaling and array fabrication.
- Oral presentation at IQUARUS, San Sebastián, Spain, July 2024, Scaling of ULTRA**RAM**™: Fabrication, characterization and simulation of III-V compound-semiconductor non-volatile memory.
- Poster at NVMTS, October 2024, Busan, South Korea, Development of scaling for ULTRA**RAM**™.
- Poster at IQUARUS, Simulations of the triple-barrier resonant-tunnelling heterostructure for ULTRA**RAM**™ memory.

Contents

Glossary of Abbreviations and Acronyms

AI	Artificial intelligence
ALD	Atomic layer deposition
ALU	Arithmetic-logic unit
BEOL	Back end of line
CFET	Complementary field-effect transistor
CMOS	Complementary metal-oxide-semiconductor
CPU	Central processing unit
CTLM	Circular transfer length method
DELTA	Depleted lean-channel transistor
DI	Deionised
DIMM	Dual in-line memory module
DRAM	Dynamic random-access memory
DUT	Device under test
EBL	E-beam lithography
ECCI	Electron channelling contrast imaging
EDS	Energy dispersive spectroscopy
EEPROM	Electrically erasable programmable read-only memory
EFO	Electronic flame off
EPROM	Erasable programmable read-only memory
EOS	Eletron optical system
FEOL	Front end of line
FET	Field-effect transistor
FeRAM	Ferroelectric random-access memory
FGMOSFET	Floating gate metal-oxide-semiconductor field-effect transistor
FIB	Focused beam ion
FN	Fowler-Nordheim
GAAFET	Gate-all-around field-effect transistor

HBM	High bandwidth memory
HCI	Hot carrier injection
IC	Integrated circuit
ICP	Inductively coupled plasma
IMC	In-memory computing
IPA	Isopropyl alcohol
LOR	Lift-off resist
LW	Laser writer
MAC	Multiply and accumulate
MBE	Molecular-beam epitaxy
ML	Machine learning
MLC	Multi-level cell
MOSFET	Metal-oxide-semiconductor field-effect transistor
MRAM	Magnetic random-access memory
MSB	Multi-scattering Büttiker
MTJ	Magnetic tunnel junction
NEGF	Non-equilibrium Green's function
NDR	Negative differential resistance
NMOS	N-type metal-oxide-semiconductor field-effect transistor
NVM	Non-volatile memory
OES	Optical emission spectroscopy
PC	Personal computer
PECVD	Plasma-enhanced chemical-vapour deposition
PLC	Penta-level cell
PMMA	Polymethyl methacrylate
PMOS	P-type metal-oxide-semiconductor field-effect transistor
PRAM	Phase change random-access memory
PROM	Programmable read-only memory
QLC	Quad-level cell

RAM	Random-access memory
ReRAM	Resistive random-access memory
RF	Radio frequency
RHEED	Reflection high-energy electron diffraction
RIE	Reactive-ion etching
ROM	Read-only memory
SEM	Scanning electron microscopy
SGT	Surrounding-gate transistor
SLC	Single-level cell
SMU	Source measure unit
SOI	Silicon on insulator
SOT	Spin-orbit torque
SRAM	Static random-access memory
SSD	Solid-state drive
STT	Spin transfer torque
TBRT	Triple-barrier resonant-tunnelling
TEM	Transmission electron microscopy
TLC	Triple-level cell
TLM	Transfer length method
TMA	Trimethylaluminium
TSV	Through-silicon via
UV	Ultraviolet
XANES	X-ray absorption near edge structure
XPS	X-ray photoelectron spectroscopy
XRF	X-ray fluorescence
1T1C	One transistor and one capacitor
1T1MTJ	One transistor and one magnetic tunnel junction
1T1R	One transistor and one resistor

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

Modern computers are based on von Neumann architecture where data are stored in memory, separately from and need to be fetched upon request by central processing unit (CPU). Memory is a fundamental requirement for computation. The $167 billion per annum memory market is dominated by dynamic random-access memory (DRAM) and flash [1]. DRAM is capable of operating at high speed and is used as the main memory in computing. Conversely, the low cost and non-volatility of flash make it highly suitable as a storage memory. However, both DRAM and flash have significant drawbacks. DRAM's volatility, need of continuous refreshing and destructive-read process are inconvenient and inefficient. The penalty paid by flash for its non-volatility is high-voltage program and erase, making it intrinsically slow and damaging to its oxide barrier, leading to poor endurance. The trade-off between non-volatility and operation speed/energy efficiency of conventional memory hinders it from being capable of meeting the ever-increasing data abundancy in the information technology era.

For these reasons, the search for a memory technology that combines the speed of DRAM with the non-volatility of flash has continued unabated for decades, resulting in the development of so-called 'emerging memories' [2]. However, even though

expected to reach $36 billion in memory market by 2030 [3], emerging memories have not been able to match DRAM in terms of speed and switching energy. On the other hand, despite technical superiority to flash, no other memory can compete with its extremely low cost/bit. Furthermore, this has led to the widely-held perception that a compromise is required between low switching energy and non-volatility. In such a scenario, a so-called 'universal memory' with a robust state that is nevertheless easily changed is "not realistic" [4].

ULTRA**RAM**™ is a compound-semiconductor, floating gate memory [5] similar to flash, where the oxide barrier which isolates the floating gate from the channel is replaced by a triple-barrier resonant-tunnelling (TBRT) structure formed by the band offsets between InAs and AlSb. Crucially, the TBRT can be switched from insulating to highly conductive by the application of a small ($\sim$ 2.5 V) voltage. This allows an extraordinary combination of characteristics for a non-volatile memory, including intrinsically-fast (sub-ns) switching speeds [6, 7], low disturb [8], high endurance [9] and ultralow logical-state switching energies [6, 9]. However, realising its full technical and commercial potential necessitates the scaling of ULTRA**RAM**™ devices. In this work, the scalable fabrication, characterisation and simulation are studied as a development of ULTRA**RAM**™ towards commercialisation. In this study, the scaling of the ULTRA**RAM**™ device was systematically investigated, covering both device design and fabrication flow. A scalable design with a feature size of less than 10 μm was proposed, representing a significant advancement towards the practical application of ULTRA**RAM**™ memory.

1.2 Synopsis

An introduction and a concise organisation of the thesis is outlined here in chapter 1 to define the scope of this work. The thesis is then followed by an introduction to memory devices in chapter 2 which covers from the essential building block of memory to final memory products. Advancements in the memory development are

also discussed, followed by the introduction and the background of ULTRA**RAM**™. In chapter 3, all facilities and techniques involved in the study, from wafer growth through device fabrication and measurement, to characterisation and simulation are listed and explored.

In the next, experimental sections focusing on ULTRA**RAM**™ fabrication are laid out in chapter 4 with analysis and discussions on three major processing designs. Then, detailed attention is given to the characterisation of the scaling work in chapter 5, including the addressing of the issues encountered in the fabrication and the memory performance, as a verification of the fabrication design. Investigation of the ULTRA**RAM**™ failure mechanism using X-ray nanoprobe technique is also analysed. Chapter 6 presents the motivation and results of the simulation study of critical interfaces of the TBRT using nextnano software. In the end, the thesis concludes with a brief summary and the discussion of future work in chapter 7. Other related details including the layout of wafers not listed in experimental chapters, the lithography pattern, etching and deposition recipes are unfolded in appendix A. Simulation basics including the parameters used for the etching modelling and nextnano simulation are briefly mentioned in appendix B.

Chapter 2

Memory Devices

2.1 Field-Effect Transistor

The field-effect transistor (FET) is a unipolar transistor that utilises an electric field to control the current through the channel. The most extensively used FET is the metal-oxide-semiconductor field-effect transistor (MOSFET).

2.1.1 Metal-Oxide-Semiconductor Field-Effect Transistor

There is no doubt that the MOSFET is one of the most life-altering inventions of the 20th century, underpinning almost every aspect of modern technology due to its diverse applications in the microprocessors, microchips and microcircuits manufactured today. The MOSFET is the fundamental element of the contemporary semiconductor industry, the basic building block of integrated circuits (ICs). In terms of memory, it's an essential component for static random-access memory (SRAM), DRAM and flash (in a modified form) memory cells, as well as for the associated driving, controlling and addressing electronics, e.g. multiplexer.

The first MOSFET-like structure was patented by Julius Edgar Lilienfeld in 1925 [10]. However, no practical device was obtained at the time due to poor material purity. In 1960, John Atalla and Dawon Kahng fabricated and demonstrated the first working MOSFET [11]. Later in 1964, MOSFETs entered the commercial market.

It has since then become the basic building block for almost all semiconductor devices, from transistor switches to advanced chips which can contain up to trillions of transistors at present. The number of transistors in a 32-GB memory card is comparable to the number of stars in the Milky Way [12]. Following Gordon Moore's prediction (known as Moore's law) that the number of transistors on a chip will double about every two years [13], numerous efforts have been centred around scaling and the solutions to tackle the issues coming with it. On the other hand, it has also attracted lots of interest in the demonstration of new structure or on the novel materials in an attempt to search for potential solutions in post-Moore era.

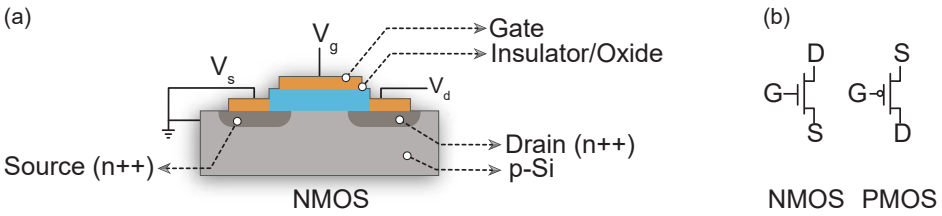

Figure 2.1: (a) Schematic of an n-type MOSFET. V_g, V_d and V_s stand for the gate bias, the drain bias and the source bias, respectively. (b) Circuit symbols for NMOS and PMOS. G, the gate; D, the drain; S, the source.

Figure 2.1 represents a generic structure of a typical n-type MOSFET (NMOS) and the circuit symbols for NMOS and p-type MOSFET (PMOS). The example given here consists of a metal gate, a gate oxide layer sandwiched between gate and channel for insulation, a channel which is an inversion layer underneath oxide that can be induced by gate control and two heavily doped contacts called source and drain which are connected to the channel when appropriate bias is applied on gate, all of which reside on top of the p-type substrate. As described in figure 2.1, an n-type MOSFET features two heavily n-doped contacts and a p-type substrate. Likewise, PMOSs possess exactly opposite doping and polarity where the majority of carriers in channel are holes. As its name stands for, the common materials exploited for MOSFET are, metal (or N^+ poly-Si) for gate electrode, oxide for insulation layer

and semiconductor for substrate which is conventionally Si. The formation of an inversion layer in the channel connecting source and drain occurs when positive charges on gate begin to attract thermally-generated electrons and repel thermally-generated holes at the same time to produce mobile electrons in the channel which balance the charge on the gate.

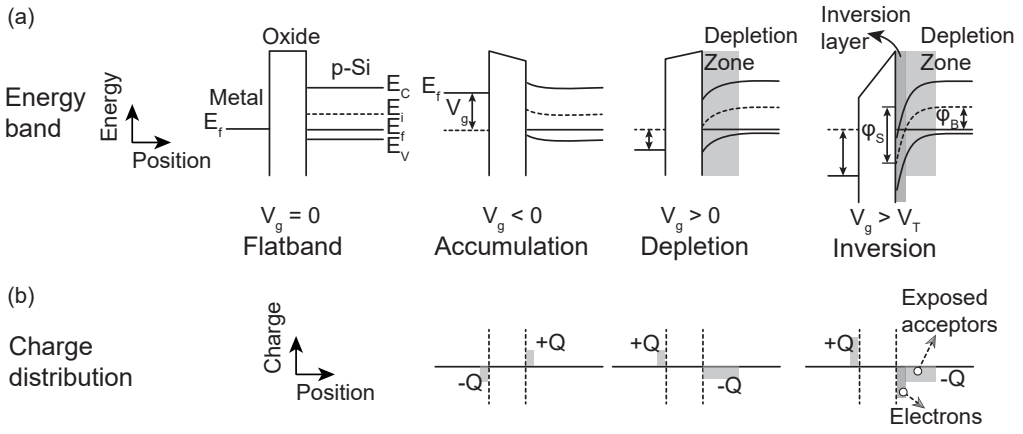

Figure 2.2: Illustration of MOSFET (a) band diagram and (b) charge distribution for different bias conditions. There is opposite band bending near surface for positive and negative conditions. Flat band, accumulation, depletion and inversion conditions are depicted. E_C, the conduction band energy; E_i, the intrinsic energy level; E_f, the Fermi level; E_V, the valence band energy; V_g, the gate voltage; V_T, the threshold voltage; Q, the charge; φ_S, the surface potential; φ_B, the bulk potential.

The working principle of MOSFETs relies on the modulation of charge concentration by a metal-oxide-semiconductor capacitance between substrate and gate. As explained by the band diagram and charge distribution at various bias conditions of an n-channel MOSFET in figure 2.2, if a voltage is applied between gate and substrate, four major scenarios can be expected: flat band, accumulation, depletion and inversion. Regarding typical conditions, the source is ground and a small fixed positive bias is applied to the drain while gate can be biased with various values. For zero gate voltage condition, a flat band is present, as both majority and minority in semiconductor are in thermal equilibrium. Therefore, there is no current between

source and drain as a reversed p-n junction. When a negative bias is applied, bands bends up and the majority carriers will accumulate at the interface, further enhancing the reversed p-n junction, so no electrons can move between source and drain. When a small positive voltage is applied, the majority carriers are repelled from the interface, bending the bands down but still no electrons can move between source and drain. When further increasing the positive voltage to higher than the threshold voltage V_T at which the device turns on, the bending continues and the intrinsic energy level E_i will cross the Fermi level, thus the minority carriers will exceed the majority at the interface, forming an electron-rich inversion layer, also with a contribution of carriers ejected from heavily doped contact regions, so that electrons can move from source to drain under the drain bias, creating a channel current. The practical criteria for inversion are considered to be the strong inversion condition where the surface potential φ_S is greater than $2\varphi_B$ (the bulk potential).

MOSFETs act like voltage-controlled resistors for varying gate voltages, but this depends on channel bias as well. Output and transfer characteristics of an n-channel enhancement (normally-off) MOSFET are plotted in figure 2.3, three distinct regions for the resistor-like behaviour MOSFET are shown in output curves. As shown in figure 2.3(a), for V_{GS} less than V_T, the MOSFET is in cut-off region where no current flows through the channel. For the gate-source voltage $V_{GS} > V_T$ and the drain-source voltage $V_{DS} < V_{GS}$ - V_T, i.e. for the linear region as indicated by the dashed line splitting two regions, the channel current increases with increasing channel bias, and it is almost linear initially. When V_{DS} reaches the value of V_{GS}-V_T, the channel current goes saturation due to what is called pinch-off. In terms of transfer characteristics, for normally-off (enhancement), the device turns on only if the threshold is reached, as shown in figure 2.3(b). For a normally-on (depletion) n-channel MOSFET, the channel current is non-zero at zero gate bias which means a negative voltage is required to fully turn off the channel. In a nutshell, the MOSFET works like an open switch when channel is not turned on where there is no inversion, and plays the part of resistor once the inversion layer

is formed. For the saturation region, the MOSFET acts like a voltage-controlled current source where I_{DS} is independent of V_{DS}.

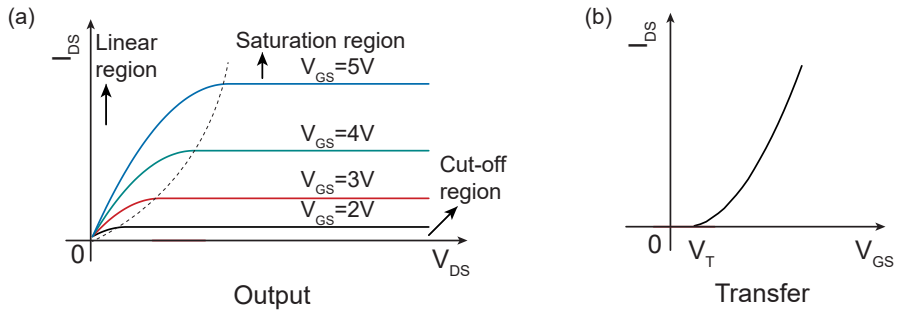

Figure 2.3: (a) Output curves at various gate voltages of an n-channel enhancement MOSFET. The dashed line denotes the boundary between the linear region and the saturation region. (b) Transfer characteristics of an n-channel enhancement MOSFET. I_{DS}, V_{DS} and V_{GS} represent the source-drain current, the source-drain voltage and the gate-source voltage, respectively.

2.1.2 Floating Gate Metal-Oxide-Semiconductor Field-Effect Transistor

The floating gate MOSFET (FGMOSFET) is a variation of MOSFET where an electronically isolated floating gate is sandwiched in dielectric between control gate and channel. It prevails in flash memory where the presence or absence of electrons on the floating gate can be interpreted to binary value zero or one. The first floating gate MOSFET was reported by Dawon Kahng and Simon Min Sze in 1967 [14] and has been widely investigated which finally led to the invention of flash. The structure of an n-channel FGMOSFET is depicted in figure 2.4. FGMOSFETs show similarity to MOSFETs, but the existence of a charge storage layer which can store and release charges creates a difference in threshold voltage of the device. The common choice for the charge storage layer in floating gate memory is polysilicon.

Figure 2.5 sketches Fowler-Nordheim (FN) tunnelling process and hot carrier

injection (HCI) approach for the operation of FGMOSFETs. The charge transport in the vertical direction through the insulator achieved by FN tunnelling process usually requires a gate voltage of around 20 V. FN tunnelling is commonly used in NAND flash. Since it takes longer for floating gate to be erased ($\sim$ 500 μs) than to be programmed ($\sim$ 200 μs), in NAND flash memories, erasing usually happens in blocks which are kilobytes and writing process takes places in pages which are hundreds of bytes. HCI is an alternative method for injecting electrons into floating gate, which is faster than FN tunnelling process and can be found in NOR flash memories.

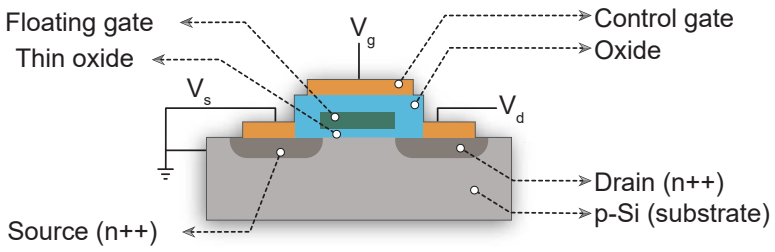

Figure 2.4: Detailed architecture of a typical FGMOSFET. V_s, the source bias; V_d, the drain bias; V_g, the gate bias.

HCI requires a higher channel voltage (around 5 V) to enable electrons in the pinch-off region gain enough energy to jump over the insulator barrier into floating gate. For HCI programming, both control gate and channel are biased, electrons travel through the channel gain sufficient energy from the lateral electric field created by the channel bias to surmount the gate oxide.

Figure 2.6 depicts the band diagram of FN tunnelling and HCI in an FGMOSFET structure. FN tunnelling is a field emission process which involves electrons tunnelling through an inclined thin barrier (triangle shape) in the presence of a strong electric field, as shown in figure 2.6(a). The current density of FN tunnelling is given by Fowler-Nordheim formula

$$J = \frac{q^3 m_{eff}}{8\pi m_{diel} h q \Phi_1} {E_{diel}}^2 exp(-\frac{4\sqrt{2m_{diel(q\Phi_1)^3}}}{3\hbar q E_{diel}}) , \tag{2.1}$$

where q is the elementary charge, m_{diel} is the effecitve electron mass in the dielectric, h is the Planck's constant,Φ_1 is the barrier height, E_{diel} is the electric field across the dielectric and $\hbar$ is the reduced Planck's constant. For the field emission where electrons tunnelling through the thicker barrier below the triangle region, a direct tunnelling can be identified by the shape of the barrier. With regard to the HCI for programming process, it's more of a thermionic-field emission which is a field emission but with the assistance of thermal process (for NOR memory cell, electrons are accelerated by the lateral electric field in the channel) for electrons to gain enough energy. This lies in the thermo-field regime, and it is a mixture of thermionic emission and field emission, as shown in figure 2.6(b). The current density for HCI in a typical MOSFET is generally given as [15]

$$J = q \int_0^\infty \nu_\perp(x, E) f_\perp(x, E) g(E) P(x, E) dE \,, \tag{2.2}$$

where q is the elementary charge, $\nu_\perp$(x, E) is the electron velocity component that is directed to oxide, $f_\perp$(x, E) is the hemi-distribution of electrons that reach the interface, g(E) is the density of states, P(x, E) is the injection probability.

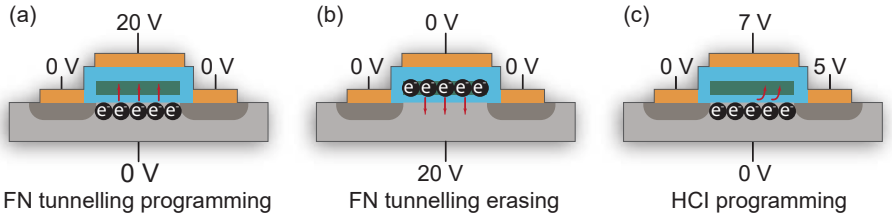

Figure 2.5: Illustration of programming and erasing processes for FGMOSFETs. (a) FN tunnelling programming. (b) FN tunnelling erasing. (c) HCI programming.

In terms of the strength of electric field, FN tunnelling typically requires a field strength to be greater than 10^8 V/m. For electric field strength lower than 10^8 V/m, thermionic emission is dominant in thermionic-field emission. FN tunnelling and HCI represent the standard approaches for programming and erasing FGMOSFETs, which also explains the limited endurance performance (10k - 100k cycles) of flash:

breakdown can be induced by electrons trapped in cumulatively generated new oxide traps under high applied field conditions in flash [16].

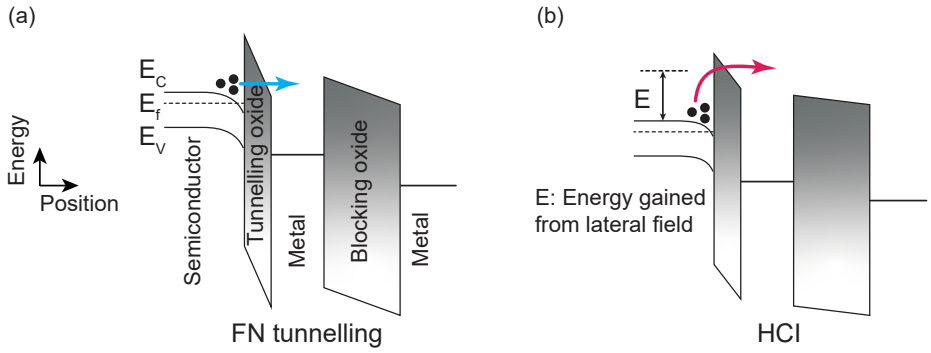

Figure 2.6: Band diagram of (a) FN tunnelling and (b) HCI in an FGMOSFET structure. The FN tunnelling is a field emission while HCI involves both field emission and thermionic emission. E_C, E_f and E_V correspond to the conduction band, the Fermi level and the valence band, respectively.

In comparison to MOSFETs, the insertion of a floating gate creates a difference of threshold voltage for programmed and erased FGMOSFET as shown in figure 2.7. The shift of the threshold voltage ΔV_T is given by

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

where Q_{FG} is the stored charge inside the floating gate and C_{FG} is the capacitance between the control gate and the floating gate, respectively. As a result, the channel shows different conductivity under erased and programmed condition as shown in the transfer characteristics curve in figure 2.7. To read the status of the FGMOSFET, a reference voltage V_{REF} halfway between V_{T1} and V_{T2} is applied. If the FGMOSFET is programmed, no current will flow, and it is erased it will. It's worth to note that, generally the floating gate is doped polysilicon in floating-gate flash, but it can also be an insulator, such as SiN [17], as a charge-trap layer for charge-trap flash. Charge-trap cell structure is dominant in 3D NAND flash [18, 19]. By storing charges in the deep-level traps in the SiN layer [20], charge-trap flash is immune to the gate leakage issue caused by the tunnelling oxide wear-out in floating-gate flash memories.

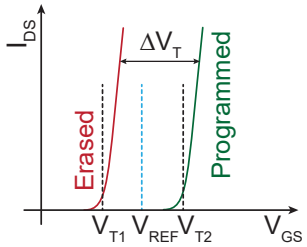

Figure 2.7: I_{DS}-V_{GS} characteristics of an FGMOSFET for programmed and erased status, showing a threshold voltage shift ΔV_T. I_{DS}, the source-drain current; V_{GS}, the gate-source voltage; V_{REF}, the reference voltage; V_{T1}, the threshold voltage for erased status; V_{T2}. the threshold voltage for programmed status.

2.1.3 Logic Gates and Memory Auxiliaries

Both random-access memory (RAM) and flash function with essential auxiliaries including amplifier and multiplexer circuits which are also based on logic gates constructed with MOSFETs.

A basic instance is the complementary metal–oxide–semiconductor (CMOS) [21] inverter which consists of a pair of FETs: NMOS and PMOS. The structure of the CMOS inverter can also be found in SRAM architecture (see section 2.2.2 below). Five logic gates, including inverter, AND, OR, NOR (NOT OR) and NAND (NOT AND), are described in figure 2.8(a)-(e). CMOS circuits for either function can be built from just six transistors, but those circuits have some undesirable features. More typically, XOR and XNOR logic gates are built from three NAND gates and two inverters, and so take 16 transistors.

Figure 2.9 shows the transmission gate structure and one example of a 4-to-1 multiplexer built with transmission gates. Starting from gates, in combination with various structures, amplifiers and multiplexers can be built. They are two important tools when it comes to dealing with numerous memory cells/bits within the chip. Amplifiers are essential for sensing small signals, which is the criteria for distinguishing binary 0 and 1 values. In order to read the stored data, the wordline of a selected cell will be raised to turn on the pass transistors, followed by a sensitive

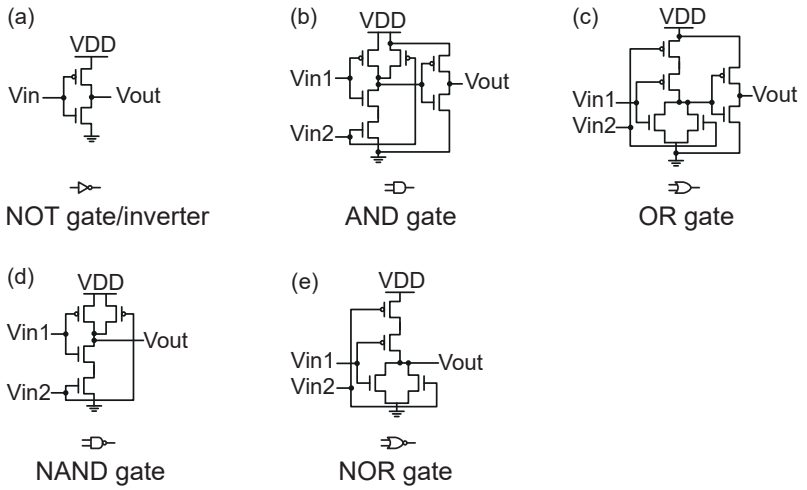

Figure 2.8: Instances of CMOS implementation in several two-input logic gates. (a) NOT gate. (b) AND gate. (c) OR gate. (d) NAND gate. (e) NOR gate. Vin, the input voltage; Vout, the output voltage; VDD, the voltage at the drain.

sense amplifier circuit to compare the voltages on bitline and reference to determine the stored binary value. Modern memories usually come with capacity of at least gigabytes or more which necessitates the multiplexers for memory output and the demultiplexers for addressing memory arrays.

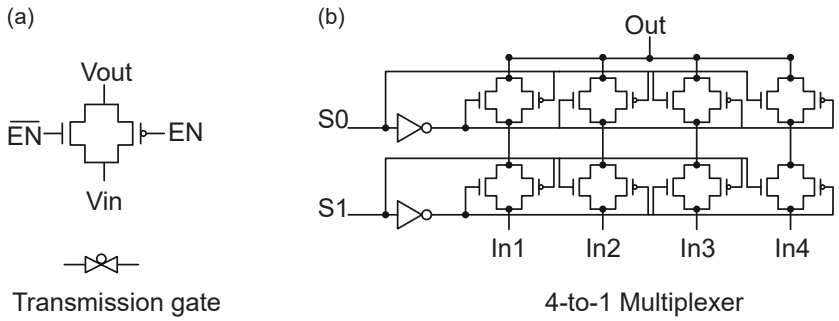

Figure 2.9: Circuit diagrams of (a) a transmission gate and (b) a 4-to-1 multiplexer. Vin, the input voltage; Vout, the output voltage; EN, enable; $\overline{\text{EN}}$, disable.

2.2 Memory Classifications

2.2.1 The Memory Hierarchy

Since the concept of the stored-program computer was introduced by John von Neumann in 1945 [22], memory has been a cornerstone of modern computers. Nowadays, a couple of memories work collectively in mainstream personal computers (PCs) and mobile devices to complete the data processing and storage task. Figure 2.10 depicts the memory hierarchy and their main usage in modern PCs. SRAM is mainly embedded in logic chips while flash and magnetic memories usually work as stand-alone storage. DRAM lies in between, it can be stand-alone or embedded in chips upon request. It's worth to note that the memory with the largest cost-effective capacity is actually provided by magnetic tape which is not listed in the figure as it is very slow and only for long-term data archiving.

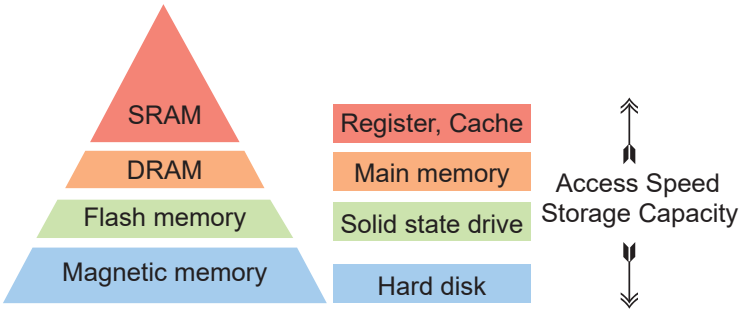

Figure 2.10: Pyramid of memory hierarchy. SRAM offers the highest speed with limited capacity while hard disk provides the largest cost-effective capacity in modern PCs.

Memories are categorized into two types in terms of non-volatility. SRAM and DRAM are volatile memories while non-volatile memories (NVMs), such as flash and magnetic memory, hold data in a non-volatile manner which can retain the stored information when the external power supply is switched off. Although SRAM and DRAM are volatile, but RAM does not have to be volatile, for instance, NOR flash is RAM but with non-volatility, and this is the whole point of emerging

memories. A fundamental difference between RAM and storage memory is that RAM requires single bit access, i.e. bits connected in parallel, whilst data storage can get away with them connected in series, which is simpler so cheaper. Although lower capacity, RAMs are capable of high-speed operations, such as running programs. Flash memory is prevalent in massive data storage due to its non-volatility and low cost. Earlier NVMs in the 1960s including read-only memory (ROM), are not categorized here, which can also hold persistent state without power supply but is not programmable. Following advanced editions of ROM including programmable ROM (PROM) [23] and erasable programmable ROM (EPROM) [24, 25], ROM made progress in ability to be programmable but still required special equipment such as ultraviolet (UV) to erase. The final evolution of ROM is electrically erasable programmable ROM (EEPROM) in the 1970s [26] which has the ability to be erased and programmed electrically, but was still restricted to low speed on byte basis and small capacity. Finally, both capacity and speed issue were addressed by the invention of flash memory in the early 1980s [27].

Type	Non-volatility	Capacity level	Readout	Read access time
SRAM	No	MB	Destructive	0.5 - 2 ns
DRAM	No	GB	Destructive	30 - 50 ns
Flash memory	Yes	GB - TB	Non-destructive	3 - 25 μs
Magnetic memory	Yes	TB	Non-destructive	5 - 10 ms

Table 2.1: Performance specifications of various memories [28–31].

Major specifications of four types of memories are listed in table 2.1. In general terms, RAM is faster and more expensive than NVMs while both flash and magnetic memory deliver higher capacity and cost-efficiency. RAMs are selected for running programs which is determined by the fast speed, while flash provides thousands of times, or even higher, capacity at affordable price due to its compact design intended

for massive data storage. As can be seen, readout for NVMs is also non-destructive which distinguishes it from RAMs which requires periodic refreshing. A limitation of the flash memory is that cycling under high-electric field can break chemical bonds in the dielectric creating leakage paths, causing failure after a few thousand cycles in modern solid-state drive (SSD).

2.2.2 Static Random-Access Memory

SRAM was invented in 1963 and a 64-bit PMOS SRAM was designed by John Schmidt in 1964 [32], marking a significant milestone for memory development, followed by the first available SRAM chip in 1969. SRAM only holds data permanently in the presence of a power supply, indicating its volatility. SRAM is one of the two basic storage elements of RAM. RAM refers to the memories for which the time to locate and access the data is not impacted by the different physical locations of the data.

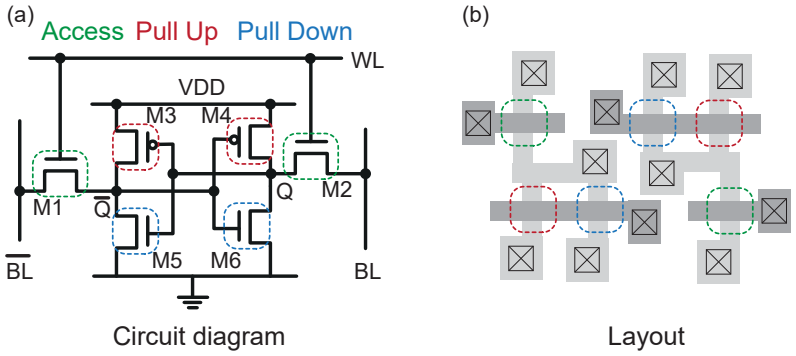

Figure 2.11: (a) Transistor configuration and (b) fabrication layout of a 6T SRAM. Data are retained by a pair of inverters in a 6T SRAM [33]. BL, the bitline; $\overline{\text{BL}}$, the opposite to the BL; VDD, the voltage at the drain. Q and $\overline{\text{Q}}$ represent binary values. The checked boxes correspond to the contacts.

An SRAM memory cell is formed by a pair of inverters in a closed loop, as shown in figure 2.11(a). A typical SRAM cell contains six MOSFETs (6T SRAM cell), each bit is stored by bistable latching circuitry. In the 6T structure, the

supply current is limited to the leakage current of the transistors in the stable state, reducing power consumption. The fabrication layout of a 6T SRAM cell is illustrated in figure 2.11(b), all rectangles with cross are the contacts, the four dark pieces are gate electrodes, of which the short ones are access transistors. Whichever cell type, no periodic refreshing if required in storing the data for static RAM, which is the distinguishable characteristic with comparison to DRAM. As an alternative, 4T memory cell structure offers a higher degree of compactness but with more power consumption. Due to its complicated architecture, i.e., the large number of transistors, and hence space, required, SRAM is mainly used for register or cache in computers with a limited capacity of MB level.

An SRAM cell has three states. To hold is to maintain its state correctly. To read is to get the data that has been requested and to write is about updating the contents. As depicted in figure 2.11, in theory, reading only requires asserting the wordline and reading the SRAM cell state by a single access transistor and bitline, e.g. M1, BL. However, bitlines are relatively long and have large parasitic capacitance. To speed up reading, a more complex process is used in practice: the read cycle is started by pre-charging both bitlines BL and $\overline{\text{BL}}$ to high (logic 1) voltage. Then, asserting the wordline WL enables both the access transistors M1 and M2, which pulls down one bitline BL voltage slightly. Then the BL and $\overline{\text{BL}}$ lines will have a small voltage difference between them. A sense amplifier will detect which line has the higher voltage and thus determine the stored value. In terms of writing, BL and $\overline{\text{BL}}$ are precharged to VDD and wordline WL is connected to VDD while two bitlines are driven to VDD and GND, respectively, to flip the state. Wordline WL is connected to GND to sustain the bi-stable operation point when SRAM is in hold status.

Figure 2.12 is a generic organisation of a static RAM. The memory cell array is the core of the SRAM, inside which, a row is activated by a global wordline and each column can be accessed individually through bitlines. Row decoder is to generate wordline signals and address line while column decoder is for selecting particular

bitlines to be connected to sense amplifiers. Further improvements can be made using an additional set of decoders to achieve reduced power and improved speed.

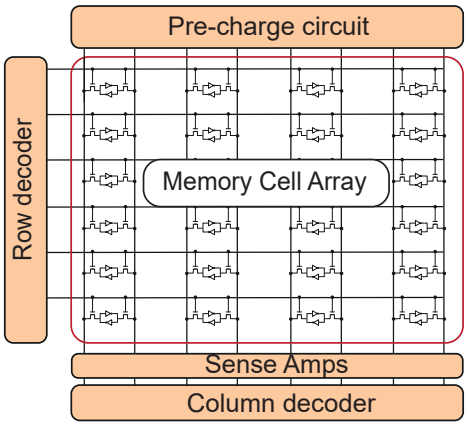

Figure 2.12: Generic organisation of SRAM. Information stored in the memory array are addressed and accessed by row and column decoders in two dimensions.

With respect to the decoder, which is an auxiliary component of SRAM, usually multiplexers are used as column decoder to select bit per column to output while demultiplexers are utilized as row decoder. Each select line can turn its transistors to a high voltage state if activated. Specifically, row decoders are used to select wordline and column decoders are used to select one or more columns for input and output of data. For 2^k lines, k address lines (selecting inputs) are required for selecting. For instance, a 16-bit multiplexer will require four select lines.

2.2.3 Dynamic Random-Access Memory

R. Dennard invented the one-transistor memory cell for DRAM in 1966 and a patent was granted for DRAM in 1968 [34, 35] DRAM is the other one of the two basic storage elements of RAM semiconductor family. DRAM requires refreshing even with power on, so it's a volatile memory.

A DRAM cell consists of a transistor and a capacitor. Due to the less complex structure which uses much fewer transistors for each unit cell in comparison to

SRAM, DRAM is much denser than SRAM. The use of a capacitor in DRAM, as shown in figure 2.13, causes the native volatility. DRAM cells lose their state over time and must be refreshed periodically, typically every 64 ms, but even more frequently (every 32 ms) in double data rate 5 synchronous DRAM. Hence, it is named dynamic RAM. Due to its higher storage density and high speed, DRAM is widely used as a main memory for enterprise, PC and mobile devices with a capacity of GB level.

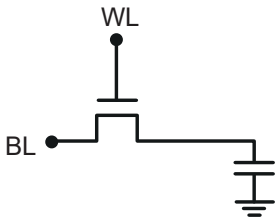

Figure 2.13: Sketch of a DRAM cell, featuring a transistor and a capacitor. WL and BL denote the wordline and the bitline, respectively.

Figure 2.14 depicts the organisation of a DRAM chip. Starting from DRAM memory array, sense amplifiers attached to column, in combination with column decoder and row decoder, a DRAM bank is built. Figure 2.15 shows the storage organisation of a DRAM with dual in-line memory module (DIMM) structure in which the array-bank-rank hierarchy is built up. All banks within the ranks share all address and control pins. Each bank operates independently of the others but can only talk to one bank at a time. This means that reading, writing and pre-charging can all be done on one bank without impacting the others. Multiple banks are used in DIMM to allow simultaneous processing on different requests. Multiple DRAM chips are used for every access to improve data transfer bandwidth. Ranks help increase the capacity on a DIMM. Due to electrical constraints, a limited number of DIMM can be attached to the bus. A functional memory is ready to communicate with CPU via memory controller and bus from this point.

With a resemblance to technology trends of logic design in semiconductor

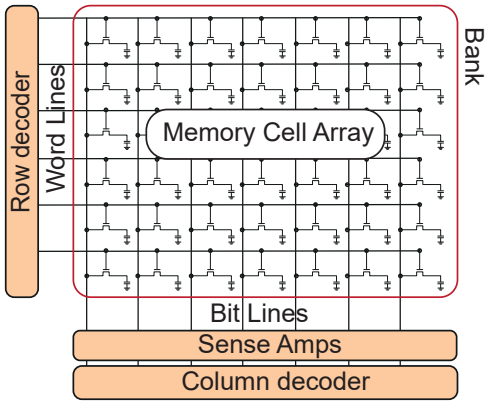

Figure 2.14: Schematic diagram of DRAM organisation. DRAM cell array forms a bank where all wordlines and bitlines are addressed and accessed by row and column decoders, respectively.

manufacturing, DRAM is making improvement by fabricating smaller devices as well. The capacity of DRAM is doubled about every two years. In terms of energy consumption, nowadays, around 25% to 40% of data centre power consumption can be attributed to DRAM system [36], making up a significant portion of energy. This can be improved by using smaller arrays, but it incurs a penalty in density and cost.

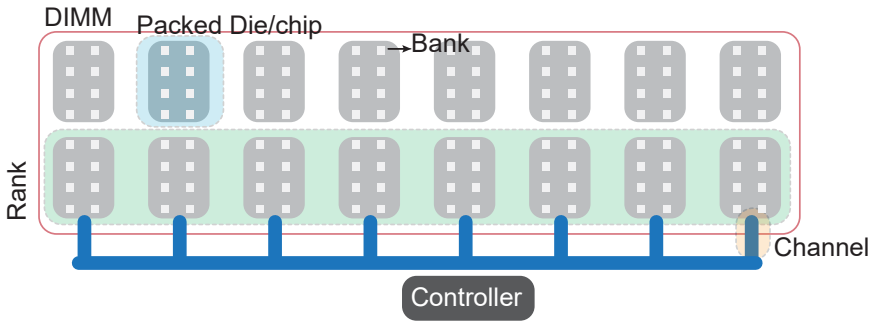

Figure 2.15: Schematic of DIMM. Array, bank, rank and DIMM form a hierarchy in the storage organisation.

2.2.4 Flash Memory

Flash is the dominant NVM, is implemented on the foundation of FGMOSFET. Following the invention of FGMOSFET, flash was first proposed by Fujio Masuoka while he worked in Toshiba and which led to the invention of flash memory [27]. In 1984 [37] and 1987 [38], NOR flash and NAND flash were presented, respectively. Flash was a successor of earlier versions of ROM such as EPROM and EEPROM at the time. Flexibility for erasing the memory to rewrite was significantly enhanced, and that is also the origin of its name, the erasure process of flash can be done in a short time. Flash memory, or more generally called floating gate memory, is a NVM that utilizes the FGMOSFET to retain the charge. Flash has achieved significant levels of operation speed, memory density and cost efficiency as a result of breakthroughs in semiconductors technologies including advanced structures in recent years.

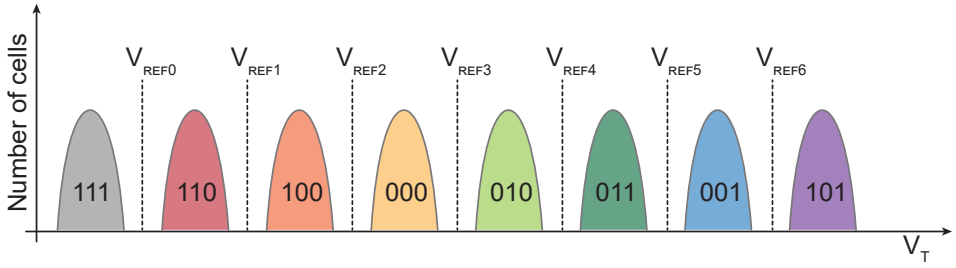

Figure 2.16: Illustration of V_T distribution in typical TLC cells and corresponding binary values. V_T, the threshold voltage; V_{REF}, the reference voltage.

The working principle of flash is based on FGMOSFET operation. In terms of data storage, conventionally, for a single-level cell (SLC) in NAND flash, the erased state with lower resistance is assigned binary 1 and the programmed state with higher resistance is assigned binary 0. Multi-levelling allows multiple bits to be stored in a single cell which significantly increases the storage density for flash. For m bits, 2^m states are required for storage. Hence, this inevitably leads to the widened threshold voltage V_T window and thus narrower margin between reference voltage

V_{REF} and adjacent V_T, which would eventually causes reduced reliability. Figure 2.16 depicts a V_T distribution of multiple cells in a triple-level cell (TLC) flash chip and the corresponding binary value assignment. For optimum read speed and least latency, the optimization of encoding for binary values assigning is required.

Cell type	Data bits	Charge states	Binary values
SLC	1	2	0, 1
MLC	2	4	00, 01, 10, 11
TLC	3	8	000, 001, 010, 011, 100, 101, 110, 111
QLC	4	16	0000, 0001, 0010, 0011, 0100, 0101, 0110, 0111, 1000, 1001, 1010, 1011, 1100, 1101, 1110, 1111

Table 2.2: Cell types and binary values.

Currently, there are four types of cell utilized in flash including SLC, multi-level cell (MLC), TCL and quad-level cell (QLC). Binary values and data bits are listed in table 2.2 for four types of flash memory cells. Regarding the bit density, penta-level cell (PLC) outperforms all these four types. A PLC NAND chip was reported with a record bit density of 23.3 Gb/mm^2 [39], and is currently undergoing active research and development.

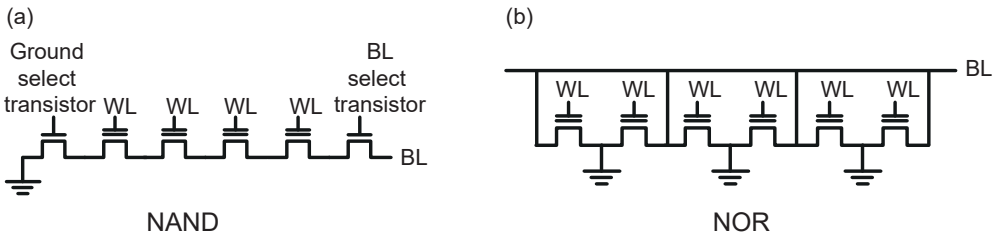

Figure 2.17: Sketch of (a) NAND and (b) NOR connections. WL, the wordline; BL, the bitline.

Depending on the ways of connection, major flash memories are either NAND or NOR wiring, being connected in a way that resembles a NAND gate and a NOR gate,

or in series and in parallel, respectively. For NAND, the bitline is pulled low only if all the wordlines are pulled high. For NOR, when one of the wordlines is brought high, the corresponding storage transistor acts to pull the output bitline low. Wiring and structure of NOR and NAND flash are shown in figure 2.17. Obviously for NOR flash, individual cells can be accessed, indicating faster access, but more silicon real estate is required, as illustrated in figure 2.18, which increases cost. Typical NAND cell size is $4F^2$ while NOR is $10F^2$, explaining why NOR flash is not used in massive storage scenarios.

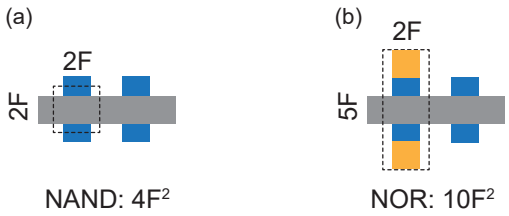

Figure 2.18: Typical cell size for (a) NAND and (b) NOR flash in manufacturing, where F stands for the dimension of gate.

Table 2.3 lists several key points concerning the comparison between NOR and NAND flash. NOR flash has the capability of direct random access due to its parallel connection with higher read speed, but still restricted to low write speed, which is determined by native slow process of the program/erase for floating gate MOSFET. NAND flash has advantages in capacity as the series connection natively enables high density design. It's worth to note that NAND flash is asymmetric in operation. As shown in the list, the smallest write/read unit is at page granularity for NAND flash, but whole block is required to be erased before a new value is written known as erase-before-write property. The minimum erasure block unit is a result of balancing the enhanced erasure bandwidth and longer erasure latency, consequently, it causes the inefficient in-place write on page where garbage collection is required. To elaborate on this, a block usually consists of thousands of cells and erasing cells in bulk is faster than individuals. In terms of writing, i.e. getting electrons into the floating gate by gradual filling from an empty cell is much more

reliable, especially for the MLC or TLC conditions. Thus, erasing before writing is a better solution.

Type	Read speed	Write speed	Smallest read	Smallest write	Random access	Capacity
NOR	Fast	Slow	Byte	Byte	Yes	Low
NAND	Slow	Slow	Page	Page	No	High

Table 2.3: Comparison between NOR and NAND flash.

Moreover, a further disadvantage of NAND over NOR is that the operation time for NAND will be even longer when the error correction code such as Hamming code to ensure data integrity is involved, which is usual for NAND while not necessary for NOR flash, a downside of the tight packing of NAND flash.

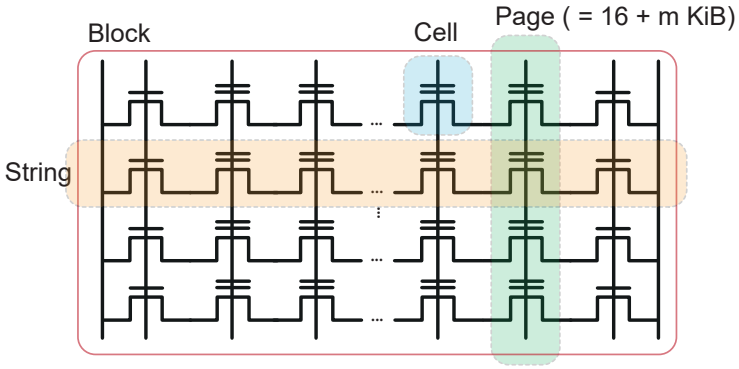

Figure 2.19: Sketch of NAND flash hierarchy. All cells connected on same bitline form a string while all cells sharing a wordline are in a page.

The hierarchical structure of NAND flash follows: cell-String-Page-Block-Plane-Die, as shown in figure 2.19 and figure 2.20. The block is made up of a matrix of strings and pages. Planes are built with multiple blocks with same bitline connections. Planes share decoders which limits the internal parallelism. There are many configurations of die to meet various design needs of original equipment manufacturers. The size of a block can be calculated by multiplying the page size

and the string number, please note that 1 byte equals 8 bits. A byte is in fact one row of the RAM memory.

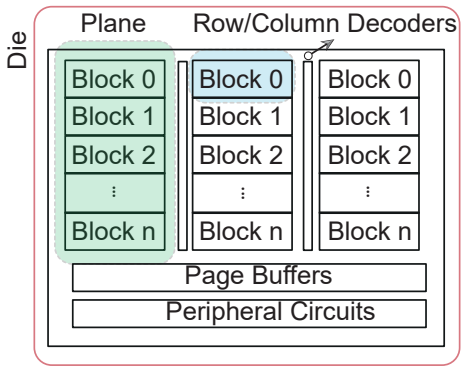

Figure 2.20: Organisation of a typical NAND die, indicating internal elements of different levels: block, plane and die.

A single die or multiple dies stacked on top of each other and bonded to the substrate and encapsulated by moulding compound then are packaged into a usable form by standard packaging such as surface mounting ball-grid array. Finally, it will be integrated onto circuit board with microcontroller addressing each die and communication interface such as serial advanced technology attachment to host will be created to make it a consumer-end product such as SSD. For a commercial product, a SSD drive also has DRAM (capacity is around 0.1% of SSD) integrated as a cache which is essential for buffering data to reduce latency.

2.2.5 Advances in Memory Development

Considerable efforts have been made, and a variety of strategies have been investigated with respect to memory development, the most important one is the miniaturisation of microelectronics [40, 41], pursing Moore's law [13] to maintain historical trend of MOSFET's scaling.

From laying the foundation to finishing touches, CMOS processing is classified in two parts: the front end of line (FEOL) [42] that comprises all the steps related

to the transistor level, and the back end of line (BEOL) [43] that encompasses the crucial sequences in semiconductor fabrication after the FEOL processes, such as secondary device integration, interconnects and packaging. One of the key points for scaling is the device structure which predominates in the FEOL block of chip manufacturing. Through memory development, the semiconductor industry has seen the rise of innovations of novel transistor architectures in the last few decades and is also witnessing the emergence of technological breakthroughs in the BEOL. Memories exploiting new materials or utilising unconventional operating principles are sprouting up with demonstration or proof of concept, diverting the attention of memory development onto a new track. Despite the expanding number of materials being actively researched, cost-per-bit and high density remain the decisive factors for new technologies to be adopted at scale by the market.

2.2.5.1 Architecture Evolution

The complexity per minimum component cost will approximately double every year, with ICs constructed of such components built on a single wafer [13]. This is Moore's law, which set the semiconductor industry on a course of developing ICs with larger numbers of smaller transistors.

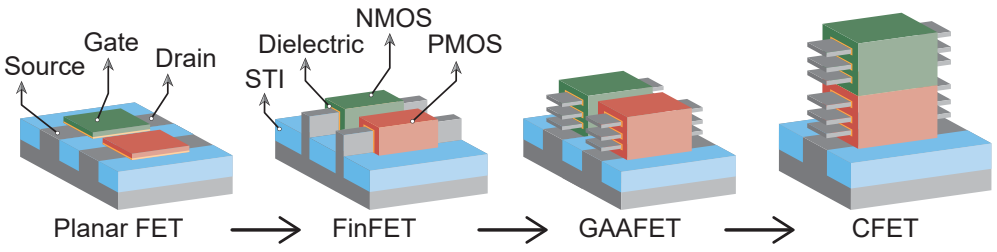

Figure 2.21: A schematic representation of the revolution of MOSFET architecture, from planar FET to FinFET, GAAFET and the latest 3D CFET.

Following the successful realisation of classic scaling scheme [44–46] in the past few decades, negative short channel effects [47] arise when the channel length is the same order of magnitude as depletion-layer widths of the source and drain junction.

This is caused by the depletion regions surrounding the source and drain extending over a large portion of the region underneath the gate. Short channel effects [48] such as velocity saturation [49], drain-induced barrier lowering [50–52] and gate-induced drain leakage [53–55] via band to band tunnelling [56] result in degradation in the subthreshold and off-state current, thus significantly reducing the gate electrostatic control of the channel. As a consequence of this collection of effects, deterioration of MOSFET performance occurs. Tremendous solutions were developed to tackle those issues arising with the very large scale integration and device shrinking, such as silicon on insulator (SOI) [57–59] and replacing conventional SiO_2 insulator with high dielectric constant material [60–62] such as HfO_2 to alleviate the parasitic capacitance and short channel effect. Meanwhile, breakthroughs in the architecture also played a pivotal role. Figure 2.21 describes the architecture revolution of MOSFET, from the earlier planar FET, fin FET (FinFET), and gate-all-around FET (GAAFET) to the latest architecture of 3D complementary FET (CFET).

Fin Field-Effect Transistor

The planar FET was introduced by Jean Hoerni while he was working Fairchild Semiconductor in 1958 [63] and paved the way for modern semiconductors since then. Dennard scaling was then proposed and experimentally verified [64], showing that reduction of the depletion layer thickness requires a proportional decrease in device dimensions (for both channel and dielectric horizontally and vertically) and applied voltage, with a proportional increase in doping. Dennard scaling worked for several decades until the IC crisis in the 2000s [65], when power and energy became the major constraints. As transistors shrink and density increases, the operation voltage is not reducing proportionately. A multiplicity of non-planar and 3D approaches were developed to prolong the scaling roadmap, of which FinFET came up as the first successful technological realisation.

Multigate MOSFET technology comes to counteract the short channel effect by taking advantage of more than one dimension. It was first mentioned in 1984 with a double gate structure predicted to solve the short channel issue [66]. The first

double gate MOSFET with FinFET-like shape was fabricated in 1989 [67] in the form of a fully-depleted lean-channel transistor (DELTA). Hisamoto later demonstrated effective suppression of the short channel effect and minimized subthreshold swing with double-gate on a Si-fin channel [68], while a new DRAM structure using DELTA was proposed in the same year. [69] Further work with gate length down to 20nm was reported in 1998 [70]. Implementation of multi-gate architectures would allow the pursuing of Moore's law until the 3 nm node [12]. Intel's shift to FinFET (or tri-gate) in production in 2011 at the 20 nm process node [71] further indicated that multi-gate technology is essential as planar MOSFET might have reached its limits, marking the commercial adoption of FinFET technology. As shown in figure 2.21, FinFET features a self-aligned gate straddling a thin silicon fin, offering faster switching times and higher current density than conventional planar architecture. FinFET has been widely deployed in memory products. Intel reported a 140 Mb SRAM for high-volume manufacturing, featuring second generation FinFET technology [72].

Gate-All-Around Field-Effect Transistor

GAAFET is the successor to the rapidly developing FinFET technology, also known as surrounding-gate transistor (SGT), offering a better short channel effect suppression than FinFET and predicted to be dominant for 3-nm-technology node and beyond [73, 74].

The first SGT with a pillar channel surrounded by a vertical gate was demonstrated in by Takato in 1988 [75]. The first GAAFET was reported in 1990 with a poly-Si gate located both above and underneath the channel, showing almost twice the transconductance than that of a conventional SOI MOSFET [76]. The first sub-5 nm GAAFET with 3-nm-fin width was reported in 2006 [77]. Later, in the same year, Singh et al. fabricated a GAAFET with a Si nanowire (diameter $\leq 5nm$) [78]. Nanosheet GAAFETs [79–81] feature multiple parallel nanosheets that share a common gate terminal as illustrated in figure 2.21. Nanosheet GAAFETs have been fabricated using conventional CMOS process [82] and higher performance over FinFET was demonstrated using 3 nm technology recently [83]. The nanosheet

GAAFET is preferred as it offers a higher drive current than nanowire-based ones [84]. GAAFETs are also promising for highly dense memory cells like SRAM [85]. From industry's view, Samsung announced gate-all-around multi-bridge channel FET [86] in mass production in 2022 [83]. Intel recently reported gate-all-around ribbon transistors with 6 nm gate length [87]. Although positive results have been reported, further effort is required to remove obstacles on way for GAAFET to fully replace the existing FinFET technology [88, 89].

Complementary Field-Effect Transistor

With the CMOS being extremely scaled, novel architectures are required to meet the demand. CFET excels with maximum device footprint reduction, while maintaining high performance [90]. As shown in figure 2.21, the CFET consists of a stacked n-type vertical sheet on top of a stack of p-type channel GAAFET or vice versa.

The first CFET-like structure was proposed as a stacked Fin-CMOS [91], stemming from the complementary nature of conventional CMOS logic design where n-FET and p-FET are controlled by the same gate. A similar structure was patented later [92]. CFET outperforms finFET according to technology computer-aided design simulation, offers structural scaling of SRAM by 50% [93] and is capable of scaling beyond 2 nm [94], presenting a promising booster for further area reductions in SRAM memories. Experimental demonstration has been reported by imec with first monolithic integration of 3D CFET, featuring functional PMOS FinFET bottom devices and NMOS nanosheet FET top devices [90]. Simulation studies also show that nanosheet-on-nanosheet stacked channels delivers superior process integration robustness compared to a nanowire-on-fin option [95]. As a leading technology, CFET architecture is currently being actively researched across different dimensions including optimization for low parasitic [96], low resistive-capacitive delay in NAND [97] and optimal configuration of nanosheet in SRAM [94].

Advances in the BEOL

With 2D geometry scaling of traditional transistors is coming to an end [98], BEOL strategies in memory technologies gain increasing attention as the memory industry proceeds to meet the unprecedented demand.

Advanced 3D IC integration [99] like high bandwidth memory (HBM) [100, 101] which takes advantage of through-silicon vias (TSVs) [102–104] to stack memory die on top of logic die enables further improvement. Emerged as a solution to address the latency in data movement between memory and CPU, HBM has been iterated for generations, and the fourth generation of HBM with maximum bandwidth of 1.65 TB/s is anticipated in 2026 [105]. With HBM integrated onto an interposer, the 2.5D system-in-package [106] delivers improved bandwidth and reduced power consumption, accelerating IC fabrication.

The vast majority of proposed 3D integration technologies are essentially 2D including TSV-based (3D by vertically stacking 2D) and interposer-based (2.5D by laterally assembling 2D), as the active part, the logic or memory unit, is constructed on the bottom of the stack due to the high temperature required for conventional FET processing [107]. Monolithic 3D integration delivers a truly 3D system by building all components directly over a previously fabricated layer in a single process flow. In 2010, world's first monolithic 3D field-programmable gate arrays by stacking 26 Mb SRAM over CMOS was reported, featuring nine layers of Cu interconnects [108]. A capacitor-less DRAM with retention more than 400 s was reported to show feasibility of 3D monolithic DRAM in 2020 [109], followed by a 3D monolithic DRAM of 1 Mb was devised recently with two stacked memory layers, operating with read and write time of 60 ns and 50 ns, respectively [110].

Monolithic 3D integration is naturally capable of ultradense integration and is projected to be promising for coming generations of applications, providing the means for both performance enhancement and power reduction that lie beyond the scope of conventional computing. Despite a growing number of research efforts on various aspects of monolithic 3D integration [111], as a cutting-edge approach, key challenges in different levels must be overcome to unleash its full potential, and

commercial monolithic 3D products do not yet exist.

2.2.5.2 Emerging Memories

A myriad of intriguing innovations have been attempted to replace some traditional memory technologies. Figure 2.22(a)-(d) highlights the cell structure for four major emerging NVMs: Phase change RAM (PRAM), ferroelectric RAM (FeRAM), resistive RAM (ReRAM) and magnetoresistive RAM (MRAM). In PRAM cells, data are stored in the phase-change material and differed by high or low resistance status. FeRAM, which is constructed from one transistor and one capacitor (1T1C) structure, retains information with its polarization in the ferroelectric layer. ReRAM cell contains one transistor and one resistor (1T1R), and the data is stored by the resistance of the resistor. With respect to spin-transfer torque MRAM (STT-MRAM), the memory cell features one transistor and one magnetic tunnel junction (1T1MTJ) structure, and the binary values are distinguished by the parallel or anti-parallel orientation of two magnetic layers in the cell. All four emerging memories are non-volatile and have non-destructive readout.

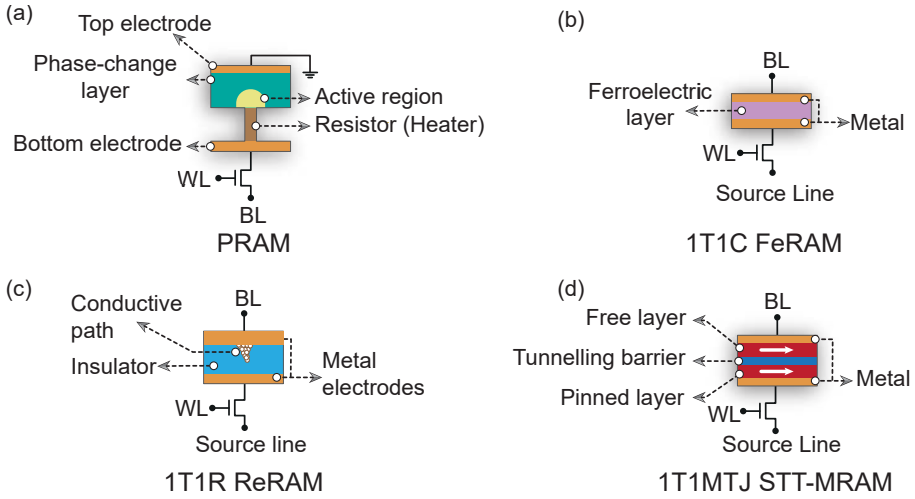

Figure 2.22: Memory cell structures for (a) PRAM, (b) FeRAM, (c) ReRAM and (d) STT-MRAM. BL and WL correspond to the bitline and the wordline, respectively.

Phase Change Random-Access Memory

Phase change memory presents low resistance when in a crystalline state and high resistance in an amorphous state. Such an order-disorder transition can be used for binary values storage.

In 1968, Stanford R. Ovshinsky observed the memory behaviour in a disordered structure of $Si_{12}Te_{48}As_{30}Ge_{10}$ where the retention of switched low-resistance state can be maintained in the absence of current and this is reversible [112]. In 1973, Ron G. Neale and John A. Aseltine reported a 256-bit array for computer memory built with phase-change material in the form of a four-element glass [113]. PRAM was projected to be scaled to 22 nm node [114], and was verified in 2011 by a PRAM cell array fabricated with 20 nm-node featuring $4F^2$ cell size [115]. Later developments were focused on reliability and stability concerning the endurance, thermal cross-talk, etc [116]. The majority of technologically useful materials for PRAM are chalcogenides such as $Ge_2Sb_2Te_5$. PRAM has regained its focus due to its potential in neuromorphic non-von Neumann computing [117–120], i.e. in-memory computing (IMC).

Albeit its advantageous scalability over flash memory and advancements in both material and technology [121–124] over the past decades, only quite a few commercial products were released, such as Micron Numonyx [125] and Intel Optane memory [126]. Challenges including the heating energy [127] and the drift of the amorphous state [128] remain in PRAM application.

Ferroelectric Random-Access Memory

Ferroelectric memory stores binary values by two stable polarization states which can be switched from one to another by applying an electric field. This is attained by the non-centrosymmetric crystal structure of ferroelectric material. For the instance with HfO_2, depending on the polarity of the externally applied electric field that is larger than coercive field, there are two possible positions of the oxygen atom in the crystal lattice, shifting up or down can be stored for '0' and '1'.

The ferroelectric phenomenon was first identified in 1920s in Rochelle salt by J. Valasek [129]. In 1963, J. L. Moll and Y. Tarui's experiment on ferroelectric FET on

CdS substrate [130] marked a milestone on the way to memory application, followed by further demonstration of memory operation with $Bi_4Ti_3O_{12}$ on a Si substrate [131]. Later attempts of the applications of ferroelectric films to realise NVMs can be traced back to 1960s, but only became feasible in the 1980s when the film deposition technology advanced to reduce the coercive voltage required to the level for computing [132]. In 2019, FeRAM with endurance $>10^{11}$ cycles,switching speed <100 ns and operating voltage 4 V was demonstrated [133]. An expanding number of materials have been demonstrated by experiment or predicted by simulation for FeRAM application including hafnium–zirconium oxide[134], inorganic ferroelectrics such as lead zirconium titanate [135, 136], barium titanate [137, 138] or strontium bismuth tantalite [139, 140], and organic ferroelectrics such as polyvinylidene flouride [141, 142]. As a contender to challenge conventional memories, FeRAM stand outs with its lowest write energy of $\sim 10^{-13}$ J per bit [143].

Notwithstanding all the progress [144, 145], apart from memories with limited capacity in MB such as EXCELON from Infineon [146] and RAMXEED from Fujitsu [147], FeRAM remains a small part of the overall memory market, especially on free-standing RAM, due to cost and difficulty of integration into current manufacturing processes.

Resistive Random-Access Memory

Resistive random-access memory (ReRAM) features a memory cell that relies on the high-resistance state and low-resistance state to store binary values. The transition is achieved by the changing of the conductive path which is usually a filament formed by either oxygen vacancies in the oxide or metallic ion in the electrode. The transition from the high-resistance state to the low-resistance state is the set operation and the reverse direction is defined as the reset operation.

Although the earlier studies on voltage-induced resistance change can be dated back to the 1960s [148, 149] on metal oxides. The ReRAM gains rapid progress and advancements in the 2000s [150, 151]. The first ReRAM array was demonstrated using 0.5 μm CMOS process in 2002 [152]. Later in 2001, a HfO_2-based ReRAM was

fabricated with an area of 10 × 10 nm^2 [153]. A TaO_x-based 2 Mb 40-nm ReRAM was reported with 10 years' retention in 2015 [154]. The BEOL compatibility of ReRAM shows its advantage for low cost ReRAM-CMOS integration [155].

Numerous efforts have been made in ReRAM research and development in the last few decades but the commercial adoption of ReRAM remains limited, with a few products available, such as ReRAM from RAMXEED [156]. Despite forthcoming challenges, ReRAM is expected to distinguish itself in the emerging wave of machine learning (ML) advancements that emphasize energy efficiency.

Magnetoresistive Random-Access Memory

Information is stored by charge in conventional semiconductor memories, while spintronics [157, 158] exploits the spin of electrons for data processing and storage. Spintronics-based memories are NVM technologies, where information is stored by the magnetization direction of magnetic layer in the cell, featuring a magnetic tunnel junction (MTJ). Two mainstream technologies, STT-MRAM and spin-orbit torque MRAM (SOT-MRAM), have been intensively researched in an effort for a fast and NVM to replace conventional DRAM technology.

The discovery of giant magnetoresistance [159, 160], the invention of magnetic tunnel junctions TMJs [161, 162], and the prediction of the spin-transfer effect (STT) [163, 164] paved the way for modern MRAM technology development. Recently, remarkable progress has been made in MgO MTJs-based STT-MRAM including demonstration of CMOS integration [165], 8 Gb STT-MRAM of perpendicular MTJ for 10 years retention [166] and Pt-based SOT-MRAM [167–170].

In terms of the industrialization of MRAM devices, STT-MRAM finds its place in niche market such as embedded memory and limited products such as STT-MRAM product from Everspin [171] and Avalanche [172], also announcements from Intel [173], Samsung [174, 175] and Global Foundries [176] proved its huge potential. In spite of ongoing advancements [177, 178] in this actively researched field, MRAM continues to encounter significant bottlenecks in the translation of basic research into microelectronic technologies.

In-memory computing

Despite considerable efforts in memory package improvement, new memory technolgies, even the optimizing memory performance by approaches such selector-only memory for compute express link [179], there is evidently increasing disparity between the speed of memory and processing units which is particularly significant in date-centric workloads such as artificial intelligence (AI). IMC offers an alternate route for higher performance by energy-efficient computation within memory [180], rather than inefficient data movement that is time and energy costly when it comes to massive data operation: the von Neumann bottleneck [181], i.e. memory wall [182, 183].

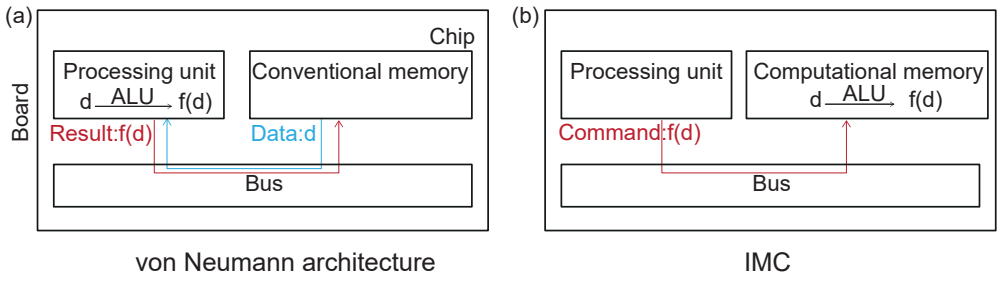

Figure 2.23: Illustrative comparison between (a) von Neumann architecture and (b) IMC where the data is processed within the memory unit. ALU stands for the arithmetic-logic unit. d, the data.

The side-by-side comparison in figure 2.23 explains the difference between von Neumann architecture and IMC structure. Using an IMC approach, much fewer data transfers are needed and it is more energy efficient. As shown, the data movement is minimized within the chip or die, eliminating the massive data traffic via the bus in a conventional architecture. Inside computational memory, certain, usually fixed, algorithms are performed with massive data, by an arithmetic-logic unit (ALU). This has huge potential for memory intensive tasks driven by AI and ML, where latency in accessing data is a key performance bottleneck. For instance, a typical 32-bit integer ALU operation consumes < 1 ns time and 0.1 pJ energy while the off-chip

memory data access costs 50 ns and > 1 nJ energy in conventional von Neumann architecture [184]. Another example is the edge computing, a convolution neural network in AI requires billions of multiply and accumulate (MAC) operations, much higher energy efficiency can be expected from IMC [185] and has been demonstrated with a tera operations per second per watt of 78.4 [186].

There has been a rising surge of interest in IMC [187], and both charge-based memory [188–190] and resistance-based memory [191–193], including the four emerging memories mentioned, have been exploited for IMC [194] and the architecture evolution has proven its advantages. However, a key disadvantage coming with the IMC is that the lack of flexibility in the algorithm, the ALU and MAC are specialized for certain workloads (optimised for AI) and become less efficient when dealing with task beyond that. Another important factor to take into account is that, as it is for MAC in AI, IMC algorithm is analogue, there is the overhead of analogue to digital and digital to analogue conversion. For instance, the ReRAM crossbar arrays implemented in IMC perform the MAC operation in an analogue manner based on Ohm's and Kirchhoff's laws [195, 196].

2.3 ULTRARAM™

A diverse range of research involving several NVMs to replace DRAM and NAND flash market were seen over the past decades. Table 2.4 summarizes the key parameters of ULTRA**RAM**™ and four viable contenders from emerging memories: PRAM, FeRAM, ReRAM and STT-MRAM.

The leading competitors in the long term for NVMs are FeRAM and MRAM that are supposed to replace NOR flash, and are commercially successful but still with limited available products in niche markets. Nevertheless, all emerging memories currently maintain a relatively small memory market presence [202]. It's perceived that there is a trade-off between information alter-ability and robust data retention, but the pursing of a so-called 'universal memory' continues unabated, this is where

ULTRA**RAM**™ makes a difference, offering higher energy efficiency but at no sacrifice of performance. The target values for endurance and switching voltage in ULTRA**RAM**™ are derived from experimental results. With further optimisation of the dielectric growth process, these parameters have the potential to be further enhanced. Retention, switching energy, and switching speed targets are based on simulations conducted at 20-nm gate dimension. In principle, these targets are attainable upon the completion of nanometre-scale integration. The cell size target is intended to inform future array design considerations.

Memories	PRAM	FeRAM	ReRAM	STT-MRAM	ULTRA**RAM**™ (target)
Endurance (cycle)	$> 10^9$	$> 10^{15}$	$> 10^6$	10^{12} - 10^{15}	$> 10^7$
Retention (year)	> 10	> 10	> 10	> 10	$> 1000^*$
Switching voltage (V)	< 4	< 3.3	< 3	< 1.5	< 2.5
Switching energy (J)	10^{-11} - 10^{-9}	10^{-13} - 10^{-12}	10^{-11} - 10^{-10}	10^{-13} - 10^{-11}	10^{-17**}
Switching speed (ns)	10 - 400	10 - 80	10 - 50	10 - 50	$< 1^{**}$
Cell size (F^2)	4 - 100	12 - 65	4 - 16	6 - 50	4-6

Table 2.4: Representative metrics of emerging memories [7, 143, 197–201]. * from the extrapolated data; ** from simulation based on 20-nm node.

ULTRA**RAM**™ works similarly as flash where data are represented by the charges in the floating gate. However, there is no oxide barrier in ULTRA**RAM**™ where the TBRT is used for charge blocking. Figure 2.24 shows the calculated band diagram vertically through the structure of the first conceptualised ULTRA**RAM**™ in the absence of gate bias [5]. As shown in figure 2.24, accumulated electrons/holes

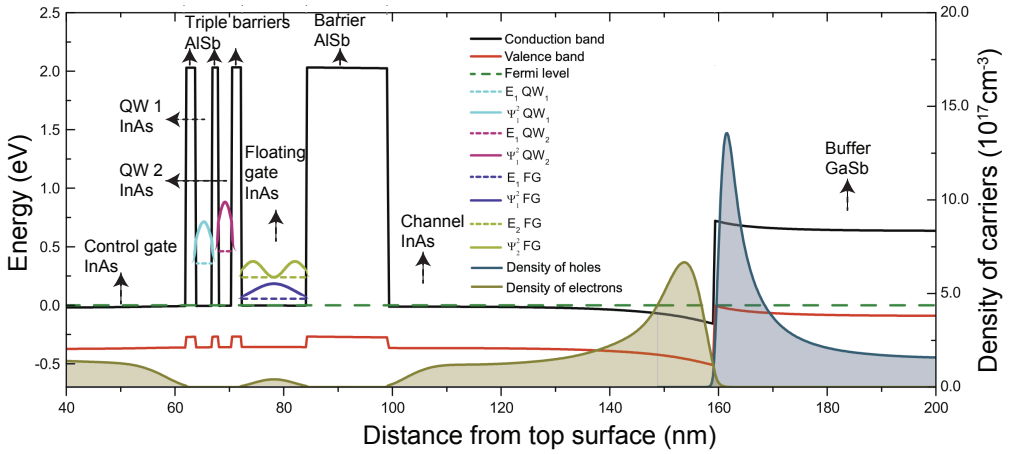

Figure 2.24: Calculated band diagram of the first conceptualised ULTRA**RAM**™ structure. E, the confined energy level; QW, the quantum well; FG, the floating gate; Ψ^2, the probability densities for the position of the electrons in the QWs. The densities of electrons and holes are plotted on the right axis [5].

are observed at the InAs/GaSb interface, but the electrons in the InAs channel are not confined to this interface, resulting in a significant electron density across the entire InAs channel. Thus, the conductance of the channel is primarily influenced by the electrons in InAs. The intrinsic InAs in the floating gate is isolated from the InAs channel by a 15-nm AlSb barrier layer, while two InAs quantum wells with triple AlSb barriers act as a resonant-tunnelling barrier between the floating gate and the InAs control gate. Therefore, in ULTRA**RAM**™ memories, the electrons stored in the floating gare are isolated by a notably large conduction-band discontinuity with AlSb. Due to the different thickness of two quantum wells, the confined energies in two InAs wells are dissimilar, and both states are higher than the control gate level. The confined levels in floating gate are also below two quantum states. With such configuration, when there is no bias applied, the passage of electrons from floating gate to control gate, or vice versa, is blocked, which means non-volatility is achieved. Similarly, when there is an appropriate gate bias applied to align the confined levels, the TBRT can be transparent for the passage of electrons, such that the erase and

write operation are achieved. The read operation is done by sensing the channel current to discern the binary states, as the presence of electrons in the floating gate depletes carriers in the underlying InAs channel, resulting the reduced conductance in the channel. With transient simulations, a switching time of less than 1 ns and a switching energy of 10^{-17} J are expected at 20 nm feature size of ULTRA**RAM**™, as a result of the combination of both small floating gate capacity and low switching voltage through TBRT [7]. In this study, due to the limitation of the lithography machine, the target of the scaling is set to 50 nm for the gate dimension, which is the smallest feature can be achieved by the e-beam lithography.

2.3.1 Multiple-Barrier Resonant-Tunnelling

Different to FN tunnelling in flash, resonant tunnelling phenomenon occurs through a resonant state in a quantum well, enhancing the tunnelling probability. The fundamental requirement is quantization by spatial confinement [203]. However, it is perceived to be difficult to realize experimentally due to structural fluctuations in both the potentials and the thicknesses [204, 205].

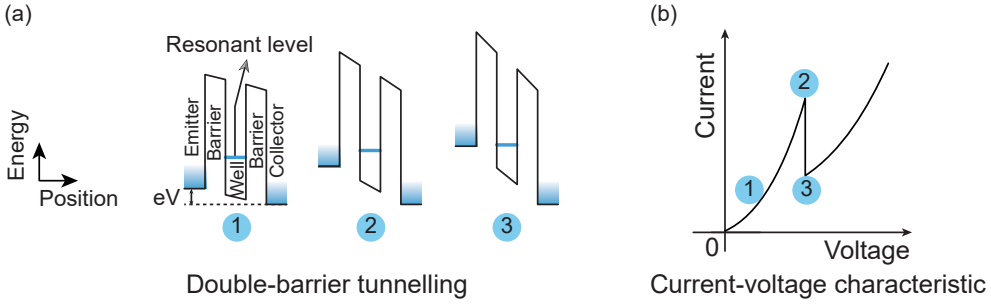

Figure 2.25: (a) Band diagram and (b) corresponding positions in I-V characteristics of a double-barrier resonant tunnelling structure. The current peak in the I-V curve occurs when the emitter level is aligned to the quasi-bound level in the quantum well.

Figure 2.25 depicts the band diagram of a resonant tunnelling structure and its I-V characteristics, whereby a negative differential resistance (NDR) can be observed.

As shown in figure 2.25(a), the band structure consists of two tunnelling barriers enclosing a quantum well. For particle in a box with infinite wall, the confined energy is given by

$$E_n = \frac{n^2\pi^2\hbar^2}{2mL^2} \tag{2.4}$$

where E_n is the energy of the nth level, n is the principle quantum number, $\hbar$ is the reduced Planck's constant, m is the mass of the particle confined in the well and L is the width of the well. Doped contacts form a Fermi sea of electrons on both ends of barriers. At a lower bias, a small current flows due to non-resonant and scattering assisted tunnelling, leakage current through surface states and thermionic emission over the barriers, as shown in the first increase in the curve. When the bias is increased, the emitter level rises relative to the resonant energy level in quantum well. Electrons incident from emitter traverses the barrier to collector side near the resonant level, current reaches maximum when the conduction band from left contact is aligned to the resonant level and the most electrons tunnel from the emission region into the resonant level in quantum well. With further higher bias, the resonant level falls below the conduction band of emitter region, thus the current will be reduced, leading to the NDR behaviour observed. At even higher bias, background effects will dominate, and the current ramps again, behaving like a conventional resistor, as shown in figure 2.25(b).

In terms of triple barrier resonant tunnelling [206], an additional barrier is introduced, forming a structure with two quantum wells sandwiched by three barriers. It can be treated as a pseudo double barrier, taking two barriers on one side as a single thicker barrier [207]. With respect to the case of asymmetric quantum wells, dissimilar discrete energy levels are generated in two wells. In principle, peak current occurs when both confined levels in quantum wells are in resonance. However, calculations show that current flows when there is alignment with just one quantum well level. Two resonances have been observed corresponding to resonant tunnelling via the ground states of the two quantum wells [208]. Figure 2.26 shows band diagram and corresponding current-voltage characteristics plot for

triple barrier resonant tunnelling with different well thickness, assuming one confined level for each quantum well. The bound state in the wider quantum well is lower than that in the narrower well. Two peaks associated to the band alignment to two bound states in the quantum wells can be seen as a variation of doublets splitting in symmetric wells [209]. Asymmetric barrier can cause charge accumulation under certain conditions [206]. For the I-V curve, tunnelling current occurs at two resonant current peaks, excess current is the smooth background that makes the valley current non-zero. For higher voltage, after the second peak, i.e. the alignment to the second confined level, thermionic current mainly contribute to the rising current.

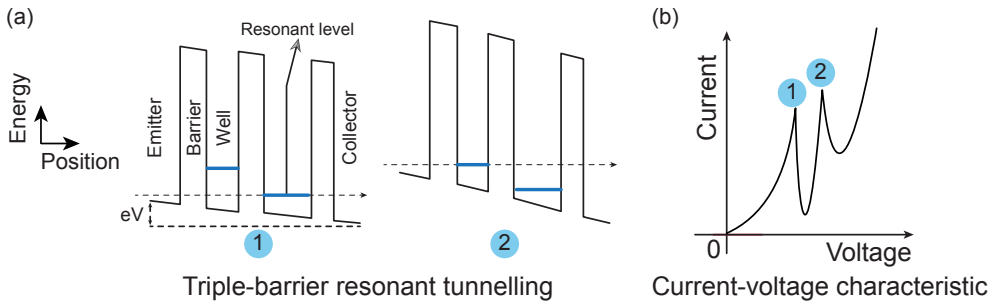

Figure 2.26: (a) Band diagram and (b) corresponding positions in I-V characteristics of a triple-barrier resonant tunnelling structure with asymmetric wells. The current peak in the I-V curve occurs when the emitter level is aligned to one of the quasi-bound levels in the quantum wells.

In resonant tunnelling, a particle can pass through a potential barrier structure with high probability when its energy matches the quantized energy level of a quantum well between the barriers. Resonant tunnelling is a specific type of quantum tunnelling where the tunnelling probability is significantly enhanced due to the alignment of the particle's energy with the energy levels in the barrier region. A particle through a single potential barrier can be described by solving the time-independent Schrödinger equation. The transmission coefficient for an electron

across a single potential barrier can be expressed as

$$T(E) = 1 + (\frac{{V_0}^2 sinh^2(\kappa a)}{4E(V_0 - E)})^{-1}, \tag{2.5}$$

where

$$\kappa = \sqrt{2m^*(V_0 - E)}/\hbar\,, \tag{2.6}$$

m^* is the effective mass of the electron, V_0 is the barrier height, E is the electron energy, a is the barrier thickness and $\hbar$ is the reduced Planck's constant. For the case of $\kappa a \gg 1$, the approximate expression is given by

$$T(E) \approx exp(-2\kappa a)\,. \tag{2.7}$$

For the complex triple barrier situation with one dimension potential, assume the total transfer matrix for the triple barrier system is M, which is built from the individual transfer matrices for barriers and wells, then [205, 210]

$$\begin{pmatrix} T \\ 0 \end{pmatrix} = M_1 M_2 M_3 M_4 M_5 \begin{pmatrix} 1 \\ R \end{pmatrix}, \tag{2.8}$$

where T and R are the transmission and reflection amplitudes. The T and R are given by

$$T = M_{11} - M_{12}M_{21}/M_{22}\,, \tag{2.9}$$

and

$$R = -M_{21}/M_{22}\,. \tag{2.10}$$

The transmission coefficient can be given by transfer matrix method by

$$T(E) = T * T\,. \tag{2.11}$$

The current density is expressed by [205, 211]

$$J = \frac{em^*\kappa_B T}{2\pi^2\hbar^3} \int T(E,V) ln(\frac{1 + exp(\mu_e - E)/\kappa_B T}{1 + exp(\mu_e - E - eV)/\kappa_B T}) dE\,, \tag{2.12}$$

where e is the electron charge, m^* is the electron effective mass, κ_B is the Boltzmann constant, T is the absolute temperature, $\hbar$ is the reduced Planck's constant, T(E, V)

is the transmission coefficient at energy E and applied voltage V, μ_e is the chemical potential in the bulk emitter, E is the electron energy, V is the applied voltage across the barriers.

In contrast to double barrier resonant tunnelling structures, TBRT offers significantly enhanced control over tunnelling dynamics. The introduction of a third barrier facilitates more precise modulation of tunnelling rates and energy levels, thereby enabling superior regulation of charge transport. Compared to their double barrier counterparts, TBRT affords an additional degree of freedom in modulating the tunnelling current, which is particularly advantageous for tailoring device performance. A better peak to valley ratio than that in the double barrier structure has been reported in TBRT [212]. Moreover, the presence of the third barrier serves as an effective blocking layer, which enhances the suppression of parasitic or unintended tunnelling currents—an essential feature for ensuring charge retention in memory applications. Furthermore, the incorporation of this additional barrier contributes to a reduction in leakage current, thereby lowering power consumption and improving the non-volatility of memory devices. The enhanced structural complexity also offers greater flexibility in engineering the tunnelling characteristics [213]. For instance, parameters such as barrier height and width, along with material properties, can be tuned to optimise the device's performance for specific applications.

2.3.2 Primary Triple-Barrier Resonant-Tunnelling Design

ULTRA**RAM**™ works on the concept of triple barrier resonant tunnelling where the unique TBRT structure is obtained by utilising the conduction band offset between InAs and AlSb. The simulations of the primary TBRT design of ULTRA**RAM**™, conducted using nextnano software, are discussed herein to provide a fundamental understanding of the device structure.

Figure 2.27(a) depicts the calculated energy band diagram of the TBRT region with standard configuration of InAs quantum wells and AlSb barriers at 300 K. The

band offset between InAs and AlSb forms the three conduction barriers i.e. triple-barrier, for the electron transport. The density of states plot shows three different confined energy levels within two quantum wells at zero bias condition. The lowest E1, then E2 and the highest E3.

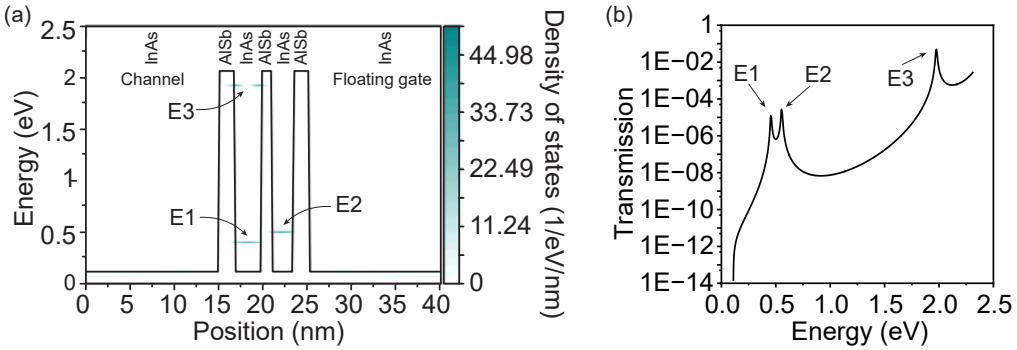

Figure 2.27: (a) Density of states as a function of position and (b) transmission as a function of energy for primary TBRT design. E1, E2, E3 are resonant levels in the quantum wells.

The discrepancy of the first two lower resonant levels in the InAs quantum wells are from the asymmetric quantum well thickness which helps to enhance the charge blocking capability of NVM. In principle, a greater degree of asymmetry enhances the charge-blocking capability at zero bias, as it suppresses resonance and thereby inhibits tunnelling. However, excessive asymmetry can lead to an undesirable increase in the program/erase voltage, as a stronger asymmetry necessitates a higher applied bias to achieve resonance. This introduces a fundamental trade-off between effective charge blocking and the required switching voltage. The minimum acceptable degree of asymmetry can be determined by considering two factors: thermal excitation and the energy broadening of the confined states. Thermal excitation is on the order of κT, while the energy broadening, denoted by Γ, reflects the resonance width of the quantised states. As illustrated in Figure 2.27(a), the confined levels E1 and E2 are broadened in energy due to elastic scattering mechanisms, including interface roughness, alloy disorder, and charged impurities.

To ensure the minimal necessary asymmetry, and assuming the resonance width is symmetric about the confined energy level E, the lowest permissible degree of asymmetry may be estimated from

$$E2 - E1 - \frac{\Gamma_1}{2} - \frac{\Gamma_2}{2} - \kappa T = 0 \tag{2.13}$$

where Γ_1 is the resonance width for the confined level in the first quantum well, Γ_2 is the resonance width for the confined level in the second quantum well, κ is the Boltzmann constant. Due to the high energy level of E3 which, in practice, is irrelevant to the tunnelling process in the memory operation, resonant tunnelling in ULTRA**RAM**™ happens only when the energy level of the injected electrons is aligned with either of the lower levels (E1 or E2), making barriers transparent for electrons to tunnel through. The tree peaks in figure 2.27(b) of the transmission curve correspond to three resonant levels in figure 2.27(a) where the energy of injected electrons matches the quasi-bound state level in the well at around 0.45 eV, 0.55 eV and 1.97 eV, respectively.

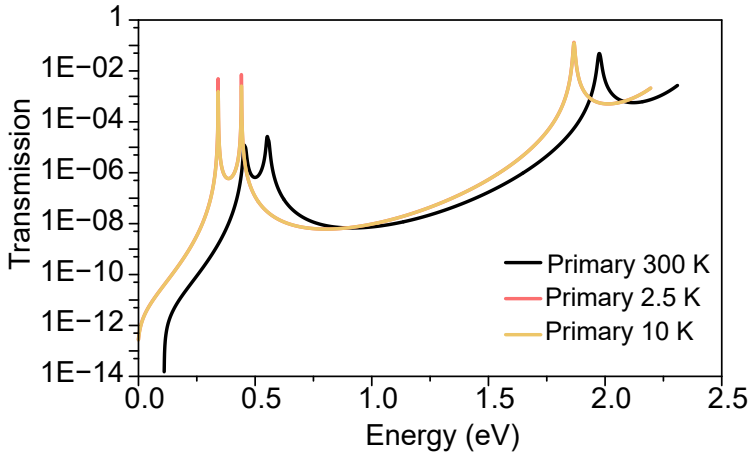

Figure 2.28: Simulated transmission of the TBRT structure as a function of energy at low temperatures. A higher transmission is observed at a lower energy for all resonant peaks.

Figure 2.28 displays the energy resolved transmission result for the same primary configuration, but at lower temperatures, to investigate the performance of the

TBRT operation. Resonant levels shift to lower energy and higher transmission coefficients are observed for both two simulated temperatures (2.5 K and 10 K) which can be credited to the lower conduction band offset of InAs/AlSb heterojunction at lower temperatures, implying the feasibility of ULTRA**RAM**™ for cryogenic applications, with lower program/erase voltages.

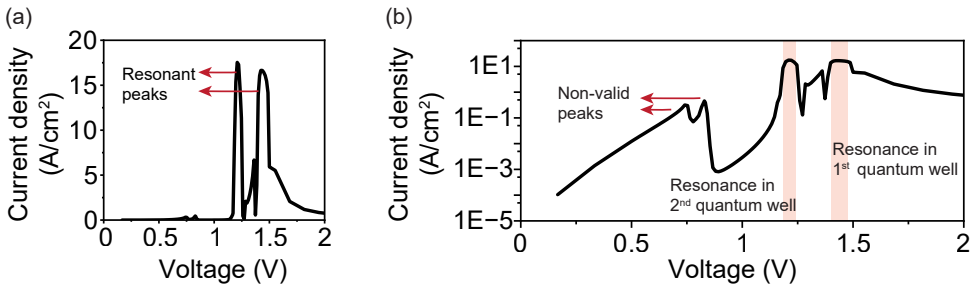

Figure 2.29: (a) Linear and (b) logarithmic plot for current-voltage characteristic for primary design of the TBRT in forward direction biasing, showing resonant peaks with NDR effect. The filled boxes represent the alignments to relevant resonant levels in the quantum wells.

To study the transport across the TBRT, a varying electric field was applied to the structure by a sweeping voltage across the two contacts. Both forward and reverse bias were simulated to examine the programming and erasing operations of ULTRA**RAM**™. Due to the asymmetric quantum well design, an asymmetric current-voltage characteristic is anticipated for the structure. From the simulation, an asymmetric J-V characteristic is clearly observed in figure 2.29 and figure 2.30, as well as the NDR features following the resonance peak in both directions at 300 K. With applied voltage, the (quasi-) Fermi level and the resonant levels change, and the applied electric field will change the conduction band profile. As a result, the transmission properties of the barriers can vary from totally blocking to electronically transparent where the resonant tunnelling occurs. The cut-off in forward bias (program) is not prominent due to the broad alignment of the resonance condition and the narrow gap between two resonant levels. Therefore, the cut-off

only produces a narrow valley in the J-V characteristic, leading to a maximum peak to valley ratio of 134, as shown in 2.29(a)-(b). In the reverse bias direction (erase), as plotted in as shown in 2.30(a)-(b)the asymmetric quantum well design provides a wider valley due to cut-off from band-gap blocking, producing a peak to valley ratio of 5690.

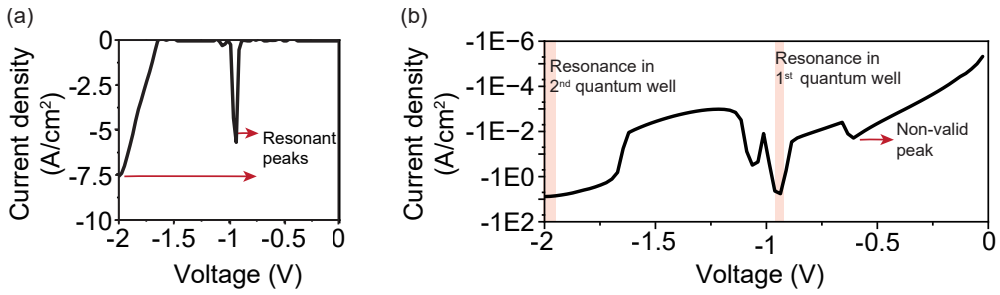

Figure 2.30: (a) Linear and (b) logarithmic plot for current-voltage characteristic for primary design of the TBRT in reverse direction biasing, showing resonant peaks with NDR effect. The filled boxes represent the alignments to relevant resonant levels in the quantum wells.

As shown in figure 2.29, the first two peaks at the lower bias seen in forward direction sweeping are generated by the ballistic transport when states from the left contact are in resonance with the quantum well resonance states. This is due to the simplified one-dimension geometry used in the simulation, and is not the circumstance for the real ULTRA**RAM**™ devices [6, 214]. Thus, all ballistic transport when the chemical potential of either contact is aligned energetically with one of the quantum well states do not contribute to the programming or erasing operations. This applies to the first peak for reverse direction as well, as shown in figure 2.30. The filled region marks the resonance peaks that contribute to the ULTRA**RAM**™ operation and corresponding locations. In the forward bias plot in figure 2.29, the peak with reduced intensity between two major resonance peaks represents the middle of the alignment transition is where the contact state is raised halfway from the lower quansi-bound to the higher one, and the contribution of

current from both resonant levels forms the transition peak.

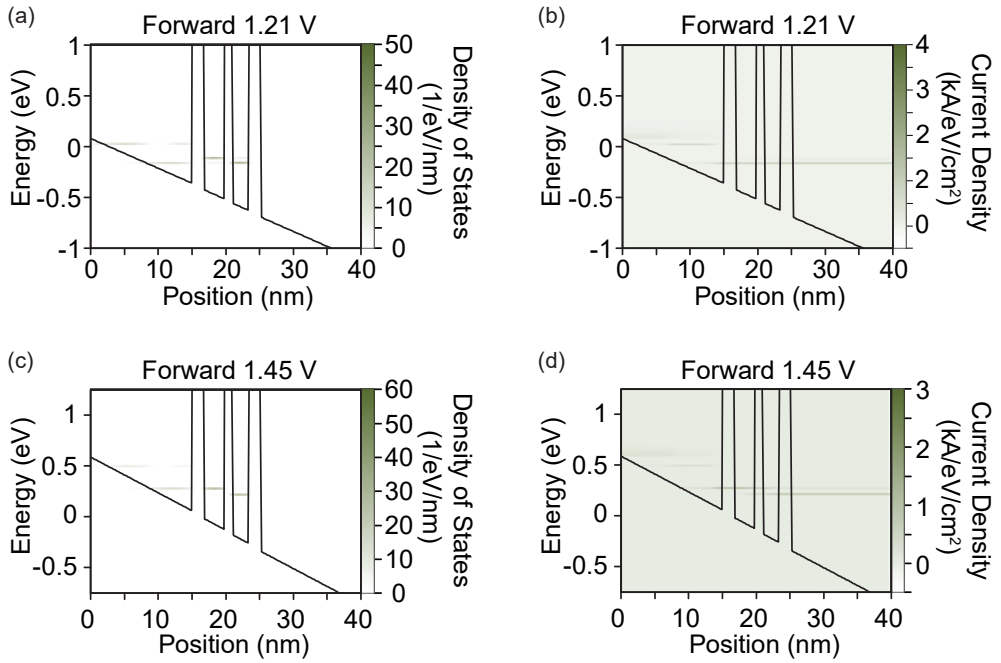

Figure 2.31: Simulation plots for two resonance conditions in forward biasing direction. (a) Density of states and (b) current density at 1.21 V. (c) Density of states and (d) current density at 1.45 V.

Figure 2.31(a)-(d) provides more details about the resonances at forward bias. In figure 2.30, the first resonance occurs at 1.21 V when the confined state in the second quantum well is aligned energetically with the triangular quantum well state in the left contact. From figure 2.31(a), one can see that the lower confined level in triangular well is aligned to the lower confined level in the second InAs quantum well. The current density in figure 2.31(b) shows the electrons injected from the left contact scatter into the triangular quantum well state in the channel and then tunnel through the triple barrier into the floating gate to the right contact. For resonance at 1.45 V in forward direction, as shown in figure 2.31(c), the scattering causes a capture of electrons in the triangular quantum well, then the alignment between the state in triangular well and the confined level in the first InAs quantum well produce the tunnelling current through the first two barriers. As seen from

figure 2.31(d), the tunnelled electrons scatter in the second quantum and lose energy before making way to floating gate. These two resonance conditions dominate the tunnelling process for ULTRA**RAM**™ memory programming operation. The current density peak between two quantum well resonance marks the transition between two major alignments during voltage sweeping.

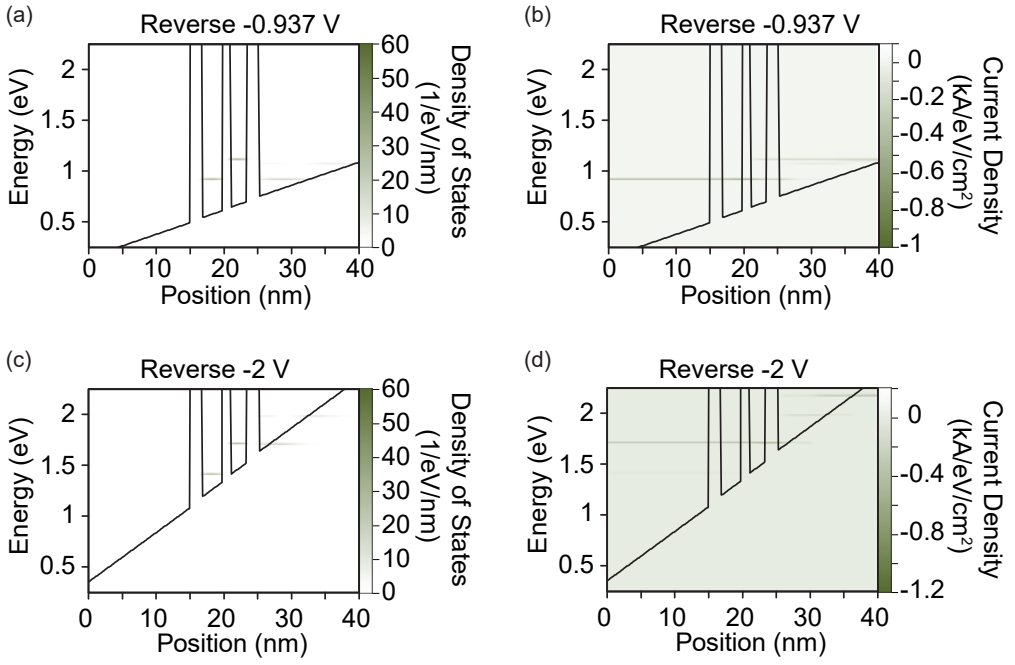

Figure 2.32: Simulation plots for two resonance conditions in reverse biasing direction. (a) Density of states and (b) current density at -0.937 V. (c) Density of states and (d) current density at -2 V.

Position resolved density of states and current density data for the TBRT in reverse bias direction are shown in figure 2.32(a)-(d) which explains the mechanism for erasing operation of ULTRA**RAM**™ memory. Figure 2.32(a)-(b) corresponds to the resonance at -0.937 V in reverse direction shown in figure 2.30, where the confined level is aligned energetically with the quantum state in the first InAs quantum well. Figure 2.32(b) shows that the tunnelling current is from the quasi-bound state in the triangular quantum well that is formed next to floating gate. In the floating gate layer, inelastic scattering causes the triangular well to be filled with electrons of

higher energies from the right contact. Electrons tunnel across the triple-barrier to form the tunnelling current. For the second resonance peak in reverse direction that occurs at a higher bias voltage (-2 V), similar scattering events cause the occupation of the triangular quasi-bound state which is aligned to the higher energy state in the second quantum well of the TBRT, producing a direct tunnelling through the triple-barrier, as shown in figure 2.32(c)-(d).

2.3.3 ULTRARAM™ Fundamentals

InAs/AlSb heterostructure and the TBRT

III-V compound semiconductors are a family of materials formed from elements of group III (B, Al, Ga, In) and group V (N, P, As, Sb) in the periodic table. An explosion of combinations of III-V elements have been produced and investigated [215–217], in binary, ternary [218], quaternary or even quinary form [219]. III-V semiconductors are found in a broad range of technologies due to their superior properties. In particular, higher mobility III-V semiconductors allow transistor to operate at higher speed than conventional doped silicon, and are potential candidates for outperforming the well-established Si CMOS [220–222].

Figure 2.33(a) depicts the band structure for InAs at 300 K, showing valleys at Γ, L and X point. The direct band gap at Γ point E_g or E_Γ is around 0.35 eV [223]. E_L at L is 1.08 eV, gap at X point is 1.37 eV and E_{SO} = 0.41 eV. InAs has a zinc blende crystal structure, as shown in figure 2.33(b), with a lattice constant of 6.06 Å. AlSb shares the same lattice structure with InAs but with a lattice constant of 6.14 Å. The indirect bandgap of AlSb is around 1.6 eV whereas the direct bandgap at Γ is 2.22 eV at 300 K. The extraordinary properties of ULTRA**RAM**™ are achieved by the TBRT which is built with alternating InAs/AlSb heterostructures [8]. As shown figure 2.34(a), the ULTRA**RAM**™ cell structure is similar to conventional flash memory, but the traditional charge blocking layer, i.e. the tunnelling oxide, is replaced by the TBRT. The unique TBRT incorporation into ULTRA**RAM**™ delivers a low switching voltage of 2.5 V (lower than flash by factors of 10) for

memory operation where the triple barrier becomes transparent to electrons. At scaled dimensions, with small capacitance, the low-switching energy E is achieved in ULTRA**RAM**™ by

$$E = \frac{1}{2}CV^2\,, \tag{2.14}$$

where C is the capacitance of the TBRT between channel and floating gate, and V is the switching voltage, respectively. Thus, the combination of low-switching energy and non-volatility are available at same time. Moreover, low-voltage operation provides advantageous endurance performance over conventional flash with more than 10^7 cycles with no degradation [9].

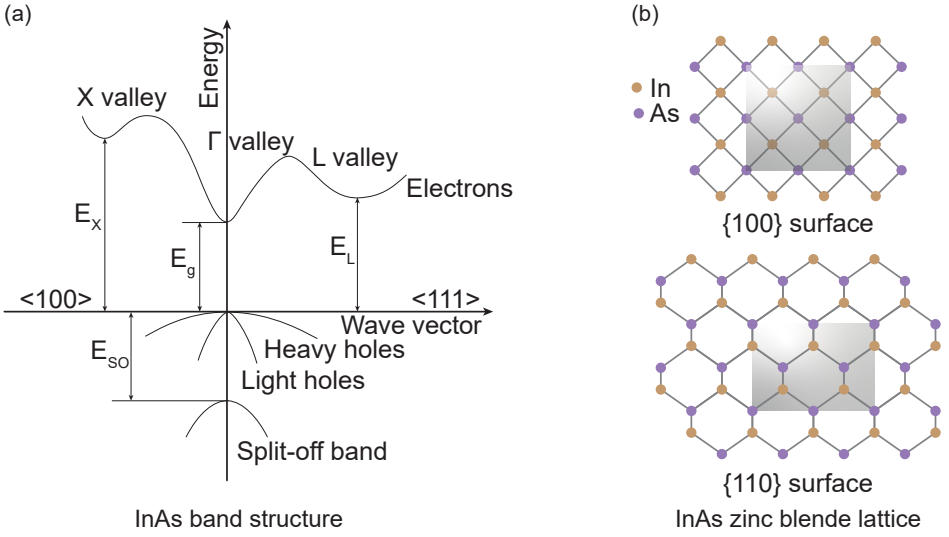

Figure 2.33: (a) Diagram of band structures of InAs and (b) its {001}, {110} surfaces of zinc blende lattice. The grey box outlines the conventional unit cell.

Within the TBRT structure, the periodically inserted AlSb layers gives a barrier height of approximately 2.1 eV [225] between InAs wells. As the vertical transport, i.e. the tunnelling, is mostly Γ like [226], without the need of phonon interaction, the conduction band offset of InAs/AlSb heterojunction is 2.1 eV which is calculated based on the direct gap at Γ point rather than the indirect band gap of AlSb. InAs and AlSb are from the 6.1 Å family [227] that has lattice around 6.1 Å, and are selected based on the band offset and similar lattice constant to minimize the

interface mismatch, as shown in figure 2.34(b). The band offset between InAs and AlSb enables the barrier construction of TBRT and the lattice constant matching makes it a pragmatic combination for growth design. The lattice constant of GaSb (6.1 Å) is between InAs and AlSb and it explains why GaSb is used underneath the InAs channel in the ULTRA**RAM**™ layer configuration, and as the (virtual) substrate.

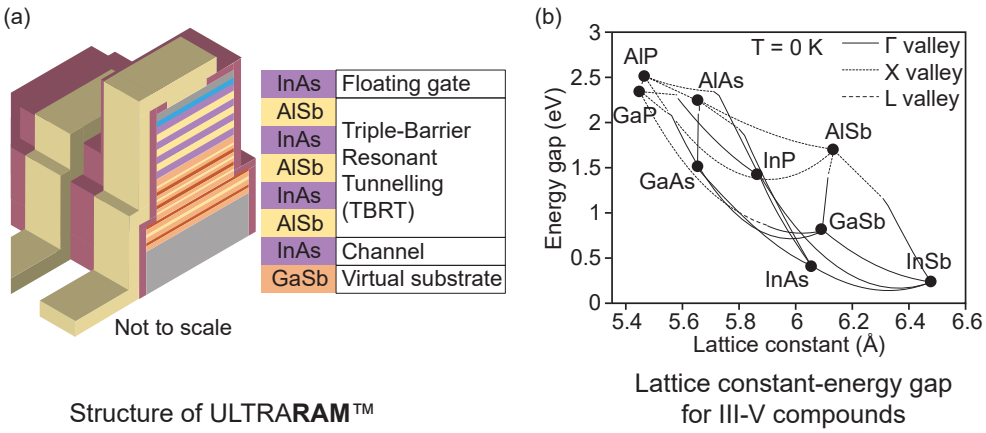

Figure 2.34: (a) Illustration of ULTRA**RAM**™ structure and (b) plot of III-V compounds energy gap as a function of lattice constant [224].

Challenges and scalability

The first prototype device of ULTRA**RAM**™ was devised with the TBRT sandwiched between the floating gate and the control gate [5], due to the gradual device degradation caused by the hole current vertically through the device, the refined layer structure was introduced [7] where the TBRT is placed between the floating gate and the channel. Given the fine structure of the TBRT, which consists of layers with thickness of down to a few nanometres, simulations concerning the layer variation during MBE growth were performed to verify the robustness of the TBRT against growth fluctuation and error [214]. However, the interface disturbance, such as alloying, which is more representative of the actual conditions observed in TEMs of the as-fabricated ULTRA**RAM**™ devices, was not considered in the simulation.

Over the long term, silicon has been proven as the basic material for ICs [13] and technology partly due to its abundance, suitable native oxide and cost. Recently, following the successful fabrication of ULTRA**RAM**™ on GaAs substrate [8], for the consideration of cost-efficiency and large-scale production, ULTRA**RAM**™ has been demonstrated on Si substrate [9] as the latest ULTRA**RAM**™ development, showing an extrapolated retention times of more than 1000 years and degradation-free endurance of over 10^7 cycles. With introduction of asymmetric program/erase voltage (due to the asymmetric wells in the TBRT), a more stable memory window of ~ 20 μA on devices with gate dimension of 20 μm was achieved during ≤ 10 ms program/erase operations, showing less current drifting. Although high performance was obtained, further improvements on epitaxy and refining of the fabrication to reduce the device-to-device variation is required. In addition, a gate leakage issue emerged as a result of the processing method employed. In particular, the demonstrated devices were fabricated using the wet-etching method which is not applicable for future scaling work.

As a viable candidate to rival conventional memories, given the competitive environment in the memory industry, it is a prerequisite that ULTRA**RAM**™ be scaled enough before its adoption to massive production. Current demonstration is limited to around 20-μm gate dimension which is far behind the industrial progress in DRAM [228] or NAND flash [229]. Also, ULTRA**RAM**™ is expected to be advantageous over DRAM once it's scaled down to nm-scale [7]. Despite all the superior properties demonstrated and predicted, a scalable design and reliable fabrication are essential to speed up the development of ULTRA**RAM**™.

Chapter 3

Research Methods

3.1 Epitaxial Growth

A molecular-beam epitaxy (MBE) system is a vacuum evaporation apparatus to grow artificially structured materials, exploring physical and chemical properties of films that do not exist in nature. These multilayer structures are prepared with alternating deposition with the state-of-the-art technique of MBE in ultra-high vacuum (10^{-8}-10^{-12} Torr) where atomically precise control of thickness and composition is achieved. Figure 3.1 illustrates a typical MBE chamber with in-situ reflection high-energy electron diffraction (RHEED).

To grow layers on a substrate, the material to be deposited is sublimated from effusion cells in the form of molecular beams incident upon a substrate that being heated and rotated. Eventually, epilayers are finished by building up these orderly layers of molecules in a sequence. Source materials are contained in effusion cells that have shutters covering them to enable the control of deposition. The heater system in these cells can be resistive filament heating using refractory material such tungsten or by electron-beam heating. For the instance of indium source material, as an example, a temperature over 700 °C is required to achieve a suitable flux which can be done by resistive filament heating.

Once the elemental molecules reach the surface, depending upon the substrate

temperature, the deposition rate and available surface energy, there are several scenarios of the deposition: physisorbed, chemisorbed or diffused to promoting growth. Three crystal growth modes can be defined in MBE growth: Frank-van der Merwe, Volmer-Weber and Stranski-Krastanov [230]. Figure 3.2(a)-(c) depicts three growth modes, where Frank-van der Merwe is more of a layer by layer mode while Volmer-Weber is island growth and Stranski-Krastanov is between them.

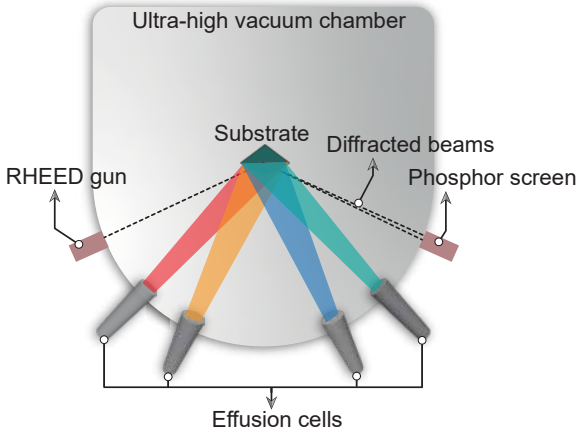

Figure 3.1: Simplified schematic of an MBE growth. Elements from effusion cells are sequentially controlled to be deposited onto the target substrate in a high vacuum condition with RHEED monitoring the real-time thickness.

Major variables such as flux rate, III-V ratio, substrate temperature and source temperature can be adjusted independently and monitored to get higher quality of growth. For instance, growing III-Sb semiconductors usually requires maintaining low substrate temperatures as antimony is readily desorbed from the surface at larger temperatures. Generally speaking, higher temperatures result in more highly ordered material but with non-sharp interface due to intermixing while lower temperatures generate more abrupt interfaces higher in quality, whilst the lower mobility of atoms introduces more point defects into the layer. In this work, the wafers used for memory fabrication were grown using a Veeco GENxplor MBE machine by Dr Peter Hodgson from the Department of Physics at Lancaster University .

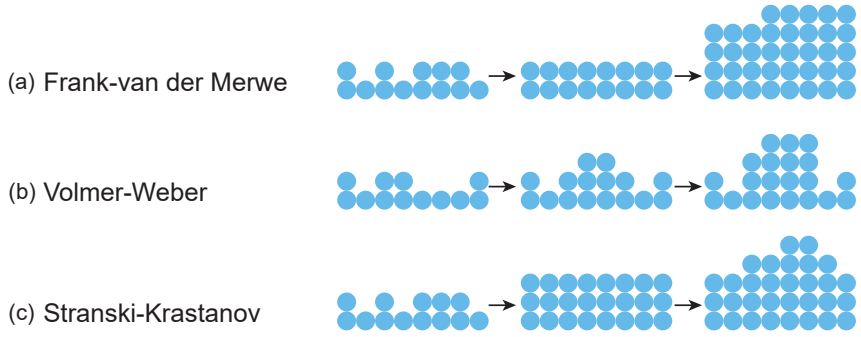

Figure 3.2: Illustration of (a) Frank-van der Merwe growth mode, (b) Volmer-Weber growth mode and (c) Stranski-Krastanov growth mode.

3.2 Fabrication

3.2.1 Optical Mask Lithography

Lithography is the foundation of semiconductor manufacturing. In a lithography process with positive photoresist, light is projected through the mask, causing the photoresist to be exposed in certain areas, the pattern is then transferred to the sample after a developing procedure in which exposed region is dissolved by developer. In a typical chip-making process, lithography will be repeated multiple times, laying patterns on top of patterns.

Optical mask lithography is the most established technology nowadays in semiconductor industry, ranging from UV, deep UV to the most advanced extreme UV technology which can generate the minimum feature size down to a few nanometres, the most feasible method for high-volume manufacturing.

Figure 3.3 shows the process of an optical mask lithography done with a mask aligner. In a standard optical mask lithography process with positive photoresist, sample is first spin-coated with photoresist that is sensitive to UV wavelengths, then the sample is mounted to the mask aligner to be in tight contact with mask that contains the pattern to be transferred onto sample, followed by the exposure to light for a few seconds during which the changes in the chemical structure of

photoresist makes the exposed region more soluble to the developer. Once the exposure is completed, the sample is put into developer to remove exposed areas, and afterwards the pattern is formed on the chip by photoresist left on the chip. Usually this is followed by subsequent steps like etching or evaporation for processing. In the context of multiple-exposure lithographic processes, alignment marks of varying scale and geometry are essential to ensure high overlay accuracy, both in terms of lateral positioning and rotational alignment. For example, crosses are common for overlay measurement and L-shaped marks offer better resistance to distortion. Such marks facilitate precise layer-to-layer registration, which is critical for maintaining structural integrity at nanometre-scale resolution. Regarding the case with negative photoresist, the exposed region becomes polymerised by light and the rest of chip is more soluble during developing process, leaving the exposed region to replicate the pattern from the mask. Photoresist, spin-coating, baking, exposure time and post-exposure developing process can affect the quality of lithography.

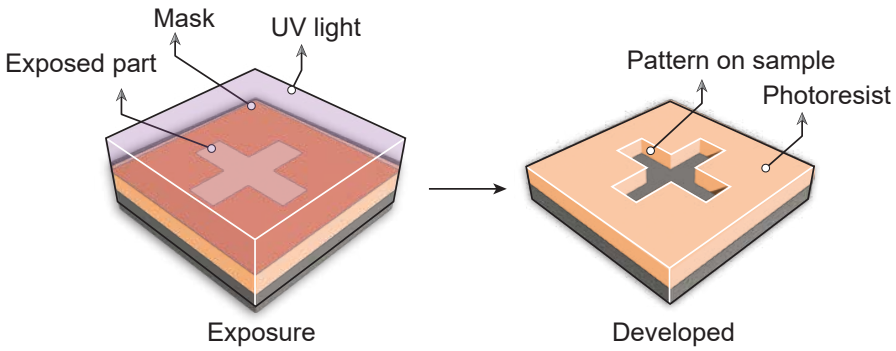

Figure 3.3: Schematic of an optical mask lithography process with mask aligner. The sample covered by photoresist is first in touch with mask and then exposed to UV light, the pattern is transferred onto the developed sample after developing process.

In this work, a SUSS MJB4 mask aligner with 365 nm and 405 nm light sources and s SUSS LabSpin6 spinner are used for UV lithography. A quartz mask is used for patterning. Photoresists including MicroChem lift-off resist (LOR) 3A lift-off

photoresist, Microposit S1813 G2 positive photoresist and MicroChem polymethyl methacrylate (PMMA) 495 A4 are used. Microposit MF-CD26 and MF-319 are used interchangeably for developing. Miccroposit Remover 1165 is used for photoresist removal or residual cleaning.

3.2.2 Laser Writer

Lithography with a mask aligner is restricted to a fixed pattern and becomes disadvantageous for developing work where rapid-prototyping and regular changes are required. The direct-write laser (LW) writer lithography process remains the same as with general lithography but uses a different method of exposure. Figure 3.4 shows how the UV lithography pattern is exposed without any mask. A moving laser spot is scanned over the pattern area to write the pattern for exposure, eliminating the use of an optical mask.

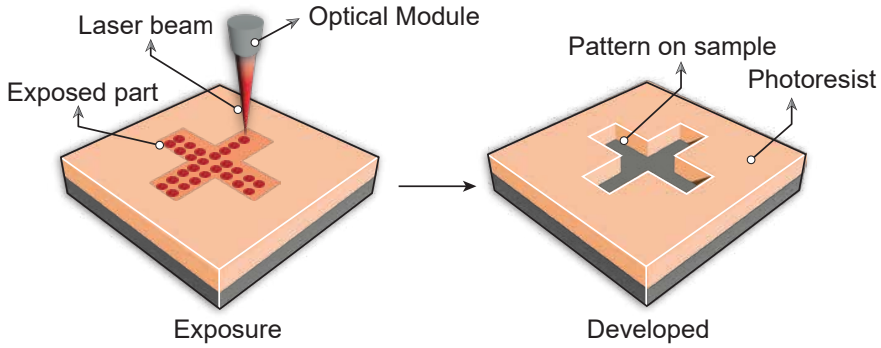

Figure 3.4: Schematic of lithography done with an LW. The pattern region to be exposed is scanned by a moving laser spot, then the pattern is transferred onto the sample after developing procedure.

Using a movable scanning laser, the LW is able to write patterns directly onto the sample with customized shapes every time, offering much more flexibility for designs with regular modifications from run to run as the pattern evolves, bypassing the long process of ordering a mask. Despite the flexibility for pattern designing, LWs are disadvantageous when it comes to large area exposures or volume production,

compared to the mask-based technique where exposure time is fixed regardless of pattern area. This can be optimized by setting up exposure parameters individually for patterns with different size and area, but is still limited to some extent.

In this work, a Raith Picomaster 100 LW with 375 nm and 405 nm lasers is employed for patterning with smaller sizes and flexible designs.

3.2.3 Reactive-Ion Etching

In semiconductor processing, the purpose of etching is the reproduction of a pattern defined by a mask in the way that involves the removal of material by etchant. Etching can be categorized into wet and dry, based on the state of etchant. Dry etching utilizes plasma as etchant, which consists of a dynamic cloud of ions, electrons and radicals. Based on the materials, a specific etching method becomes a balance between physical and chemical mechanisms.

In the plasma-etching-system configuration of reactive ion etching (RIE),the wafer is placed on a radio frequency (RF) powered electrode which is the bottom plate. Since the ions are accelerated by the bias towards the wafer, RIE etching is via the ions/radicals generated in the etching gas near the biased substrate surface. It can be physical etching when inert gas only is used, or a combination of physical and chemical etching when reactive gases are employed.

Figure 3.5 illustrate the process for reproducing the pattern defined by a mask using RIE. Firstly, the etchant gases are injected into chamber which is followed by a stabilisation stage. Oxygen plays a critical role in nearly all etching recipes, serving as an essential gas in a wide range of etchings. Secondly. Once the pressure and flow are stable, the plasma strikes by appropriate ignition. The region not protected by photoresist is then exposed to plasma and etched away. Based on the etching rate, the process is stopped at a certain time, after the pattern from the photoresist is transferred onto sample surface eventually.

In this work, an Oxford Instruments Plasma Pro NGP80 RIE is used for all RIE dry etching. CF_4, CHF_3, O_2, Ar and SF_6 are included in the gas supply.

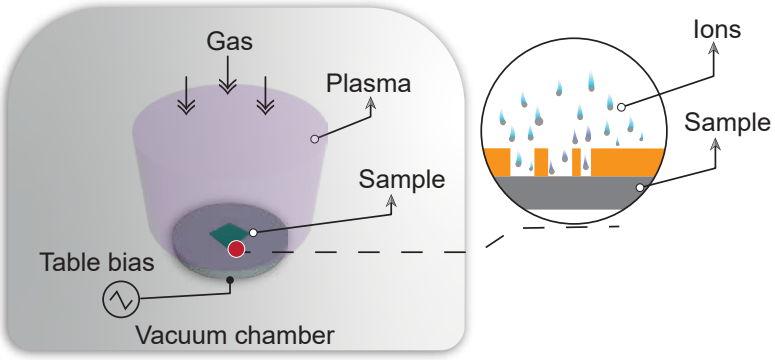

Figure 3.5: Illustration of RIE etching. The reactive ions are dragged onto the substrate by table bias to sputter or chemically react with the uncoated regions on the sample to reproduce the pattern from lithography.

3.2.4 Inductively Coupled Plasma Etching

Inductively coupled plasma (ICP) etching is a widely used dry etching method in semiconductor processing. In the configuration of RIE, there is only a single RF power supply making it difficult to shift the process through a range of chemical to physical mechanisms, and the plasma created in such a system is a low-density plasma.

The fundamental process for creating ions and radicals is the collision with accelerated electrons. To address the lack of plasma, using magnetics increases the electron's path to cause more collisions is an alternative approach to simply adding more power and gas. A variation of this approach to overcome low plasma density is ICP where a second RF power supply for wafer bias is required to control the ion flux. ICP is a non-capacitive method to generate plasma. As shown in figure 3.6, a coil is wrapped around the cylinder and separated from the chamber via a dielectric window. The inductor functions as magnet in terms of creation of a magnetic field and the oscillating magnetic field induces an electric field, and, as a result, the electrons move in spiral paths and a higher density of plasma is generated. With the configuration of two independent RF power supplies, a range of processes

that span purely physical etching to purely chemical etching can be achieved in ICP.

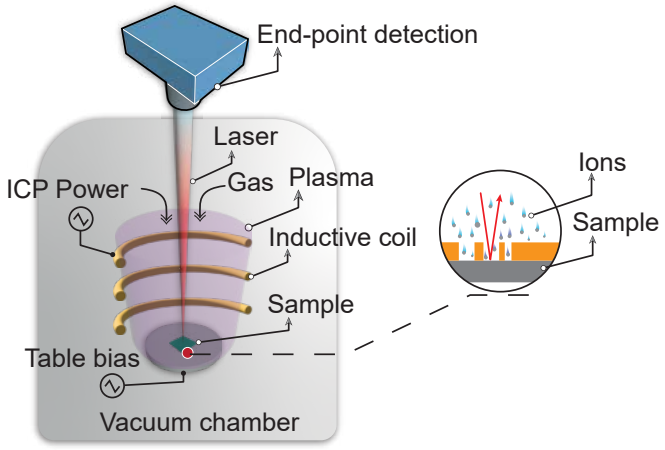

Figure 3.6: Illustration of ICP etching with end-point detection technique. The introduction of the coil contributes to higher etch rate. The laser interferometry shines and collects reflected light to make real-time etching control feasible.

Etching gases, flow, RF (bias) power, ICP (source) power, temperature and chamber pressure are factors to be considered in terms of a specific etching, which affects the etching from all aspects including etching rate, sidewall profile, selectivity, surface roughness etc. Owing to the high-energy nature of ICP etching, particular attention must be given to defect formation, surface roughness, and precise control of etch thickness. With appropriate optimization, isotropic, directional and vertical etches can be obtained. With regard to the III-V etching using BCl_3 recipe, the etching products are all $GaCl_x$ Al_x and so on. Most of the etching products are volatile.

In this work, an Oxford Instruments Plasmalab System 100 ICP is used for ICP dry etchings. Gas supplies of BCl_3, Cl_2, CH_4, H_2, SF_6, Ar and O_2 are provided with the instrument.

End-Point Detection Technique

Device performance depends on the quality of etching, especially for those with fine features and thin layers, imposing a significant demand on etching control.

Figure 3.6 includes an end-point detection kit atop the chamber, an optical in-situ metrology tool. Optical emission spectroscopy (OES) and laser interferometry (single or multiple wavelengths) are mainly equipped in dry etching systems.

OES detects specific optical emission signals from the etching plasma, more specifically, from the etching by-products, to determine the etching transition from layer to layer. For instance, in Si_3N_4 etching, depending on the etching recipe, optical emissions from N_2 (337 nm), CN (387 nm) and N (674 nm) can be used for end-point detection. As OES basically measures an averaged signal from whole chamber, a relatively large sample and faster etching rate are favoured for effective monitoring. OES offers useful information of the etching condition, this is particularly helpful for chamber cleaning, a essential and standard process for all dry etchings. On the other hand, since it detects emission from particular atoms or molecular species, it's not sensitive to in-layer thickness but the transition. Therefore, the OES technique is mostly used for end-point detection of layer transition.

Light interferometry detection uses interferometric process for precise and reliable control of etching on-the-fly. It works on the interference principle which occurs when monochromatic light hits the sample surface and different optical paths are created due to film thickness variation. This allows in-layer thickness monitoring in real time, providing enhanced process control compared to OES which can be used for layer transition monitoring only. Since it focuses on the layer only, no large area is required and the selection of light spot needs to be in an unmasked region, so manual positioning and focusing are performed before each run.

Real-time optical feedback from reflectance allows making the decision to terminate the etching once a target layer has been removed. For example, etching from high-reflectivity metal such as aluminium to lower reflectivity material will give trace a drop, indicating the full removal of aluminium. For a transparent layer etching, a sinusoidal signal is generated by optical wave interference, providing multiple points (ripples) to end the etching accurately as desired. In this case, the

etching depth d for one period is calculated to be

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

where λ is the wavelength and n is the refractive index of the material. For more complicated multilayer structures grown by MBE, modelling is anticipated for the reflectance trace analysis to guide the etching control.

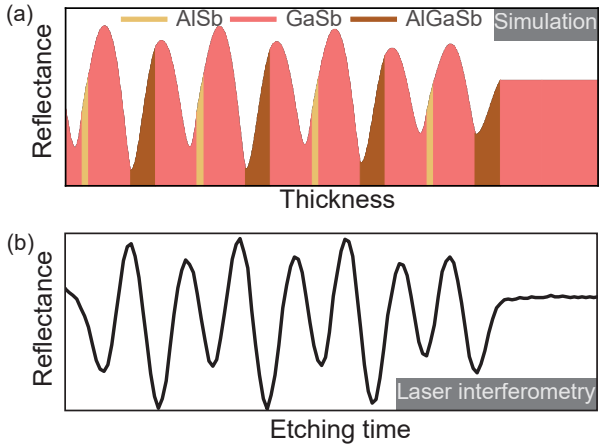

Figure 3.7: Comparison between (a) simulated reflectances and (b) data obtained from laser interferometry of a periodic GaSb/AlSb/GaSb/AlGaSb structure. The simulation is done at 670 nm, the same wavelength used in the laser interferometry: the two curves show good matching in terms of the number of peaks and general shape.

Figure 3.7 shows a good matching of etchings between simulation and from laser interferometry at 670 nm wavelength. Depending on variation of recipes, the broadness of the peak might be different due to dissimilar etching rate from material to material. As shown in figure 3.6, a laser at a specific wavelength impinges on the surface being etched, which reflects the ray into a detector for signal acquisition. The mechanism for analysing the reflectance is based on the transfer-matrix method used for electromagnetic waves through a stratified medium or in our case, a stack of MBE-grown epilayers. In a layered structure, the reflections are partially transmitted and then partially reflected. As the path length changes, these

reflections can interfere constructively or destructively. Consequently, the overall reflection of a layered structure is the sum of a number of reflections. Consequently, the reflectance will change as the etching proceeds, enabling the endpoint monitoring of the etching process. In a simple case, etching films that are transparent at the laser monitor wavelength produces a sinusoidal signal due to optical wave interference, which offers multiple stopping points to end the etch upon.

In this work, interferometry with 670 nm laser from Horiba and a multiple wavelength interferometry with 340 nm, 365 nm & 405 nm from LayTec are used for etching end-point detection. Simulations of reflectance for various structure are performed using SimEtch [231] and open source package EMpy [232].

3.2.5 Atomic Layer Deposition

Dielectric deposition involves creating a film with insulating properties onto a substrate or sample, providing gate dielectric function or as a passivation layer.

Atomic layer deposition (ALD) films are grown onto a substrate by a self-limiting sequentially-alternating, injected gaseous precursors. Figure 3.8 explains the sequential ALD reaction. By defining the precise number of cycles run by ALD, film thickness can be well controlled down to atomic scale. ALD is widely used for dielectric growth of Al_2O_3, HfO_2 and SiO_2.

Figure 3.8 depicts one cycle of a typical alumina growth. Trimethylaluminium (TMA or $Al(CH_3)_3$) and H_2O are the two precursors used. Depending on the substrate, alumina growth can be done at various temperatures, ranging from 80 °C [233] to 180 °C or higher [234], impacting the crystallinity, structure (α-Al_2O_3 or others), breakdown field and chemical stability. The process starts with injected TMA reacting with H_2O absorbed onto the surface, generating CH_4 as a product. In the following step, all dangling H bonds from the absorbed water are completely replaced and occupied by TMA, marking the termination of the self-limiting reaction. The excess precursor and products are evacuated by purging. Then, the water precursor is pulsed into chamber to react with the exposed CH_3

from the previous step. This is also self-limiting, and it stops once all the CH_3 is depleted by H_2O. Eventually, all the excess molecules including reaction products are purged with inert gas to leave the H bonds on the surface as in step 1, marking the start for next repeated process. After certain cycles, the alumina film is achieved with target thickness. The total reaction of the process is

$$Al(CH_3)_3 + \frac{3}{2}H_2O \rightarrow \frac{1}{2}Al_2O_3 + 3CH_4\,. \tag{3.2}$$

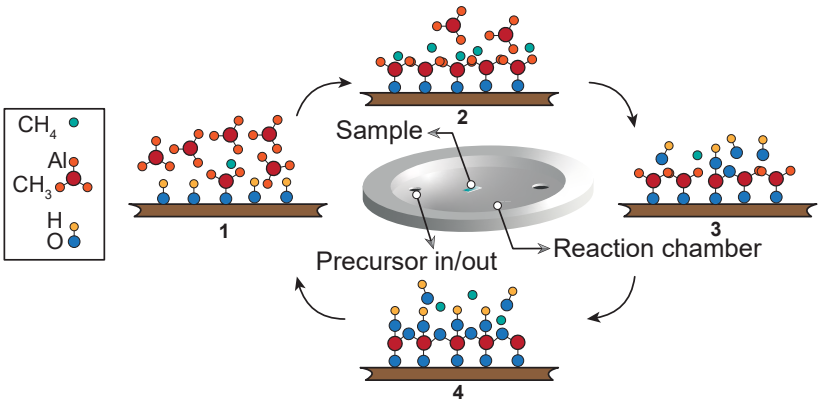

Figure 3.8: Schematic representation of the sequential ALD process for Al_2O_3 growth. In step 1, the $Al(CH_3)_3$ precursor is pumped into the chamber and absorbed by hydroxyl on the surface. The reaction produces CH_4. This is a self-limiting reaction as the precursor does not react with absorbed Al species. In step 2, the reaction products and un-reacted precursor are removed from the chamber by flowing inert gas to prepare the top surface for next process. In step 3, H_2O is introduced to the deposited methyl surface. The reaction creates the Al-O-Al bridge and leaves new hydroxyl on surface. CH_4 is released as a by-product. In step 4, a same removal process for reaction products and un-reacted precursor. Repeat the four steps again to grow an alumina layer with desired thickness. Due to the self-limiting process, only one layer of alumina is grown after each cycle.

In terms of the growth of HfO_2 and SiO_2, precursor tetrakis (dimethylamino) hafnium with H_2O and bis (diethylamino) silane with ozone (O_3) are used, respectively.

In this work, a Veeco Savannah S100 ALD is used for alumina deposition. The precursors supply of TMA, tetrakis (dimethylamino) hafnium, bis (diethylamino) silane, O_3 and H_2O are included in the machine.

3.2.6 Plasma-Enhanced Chemical-Vapour Deposition

The plasma-enhanced chemical-vapour deposition (PECVD) technique is popularly used for silicon dioxide (SiO_2) and silicon nitride (Si_3N_4) conformal film deposition, delivering a higher growth rate than ALD with increased density of plasma.

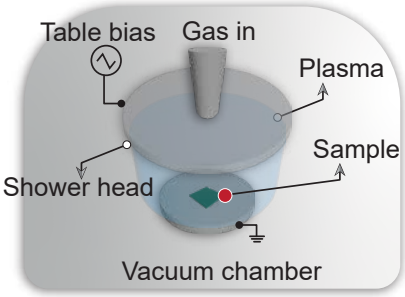

Figure 3.9: Illustration of a PECVD chamber. Reactant gases are injected from an engineered shower head and then react near the substrate to produce thin films.

Silane (SiH_4) and nitrous oxide (N_2O) are used to grow SiO_2, while a combination of SiH_4 and nitrogen (N_2) are used to deposit Si_3N_4. Figure 3.9 shows the structure of a PECVD chamber. Similar to RIE and ICP, PECVD deposition is accomplished in a vacuum chamber which requires a turbo pump and backing pump in place for maintaining proper pressure. The chemistry for Si_3N_4 is expressed in a simplified version by

$$3SiH_4 + 2N_2 \rightarrow Si_3N_4 + 6H_2\,. \tag{3.3}$$

Si_3N_4 can also be synthesized using SiH_4 with NH_3 as reactant gases, which is a more complex mechanism and might result in more H in the film [235]. A typical deposition process involves pumping, gas stabilisation, strike, deposition, purging and venting to atmosphere. Pressure, gas ratio, flow and temperature are parameters

to be considered in regard to the deposition rate, thickness and film quality.

In this work, an Oxford Instruments Plasma Pro NGP80 ICP CVD (PECVD) is used for silicon nitride deposition. The supply gases of NH_3, SiH_4, N_2O, N_2 and SF_6 are included in the instrument.

3.2.7 Thermal Evaporation

Metallisation is about depositing metal film onto devices for device contacts or interconnection. Common techniques include thermal evaporation, sputtering, e-beam evaporation, induction evaporation etc. The selection of material for metallisation also has an impact on device performance as the metal-semiconductor interface plays an important role in determining the electrical properties of the device contact.

Figure 3.10(a)-(b) shows accumulation and depletion type contacts, and the case where the contact barrier is formed due to Fermi level pinning. As the two materials are brought together, to match the Fermi level, the bands in the semiconductor bend, which subsequently form the metal-semiconductor interface. In figure 3.10(a), the metal-semiconductor interface is ohmic (or low Schottky barrier height) when the metal work function Φ_{m} is smaller than the semiconductor work function Φ_{s}. In figure 3.10(b), a Schottky barrier (rectifying) is formed when $\Phi_{\mathrm{m}} > \Phi_{\mathrm{s}}$. The Schottky barrier Φ_{b} can be calculated from the electron affinity X by

$$\Phi_b = \Phi_m - X\,, \tag{3.4}$$

based on the Schottky–Mott rule of Schottky barrier formation. In the case of Fermi level pinning as shown in figure 3.10(c), the band bending in the semiconductor is independent of the metal work function. Multiple models have been proposed for explaining the Fermi level pinning [236], such as metal-induced gap states model which attribute the phenomenon to the penetration of metal wave function into semiconductor. At the interface, the infinite periodic potential assumption in the bulk is violated, localized energy states within the forbidden band are allowed. After

the metal and the semiconductor are brought into contact, the wave functions from the metal penetrate and fill gap states in the semiconductor. Thus, they are named metal-induced gap states. The presence of surface states causes the Fermi level to be pinned at a fixed position near the mid-gap of the semiconductor, regardless of metal work function.

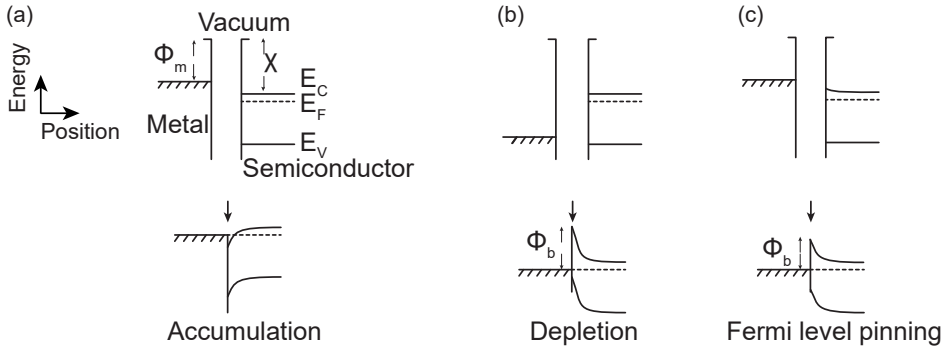

Figure 3.10: Band diagrams of three contact types with instance of an interface between a metal and an n-type semiconductor. (a) Accumulation. No contact barrier is formed at the interface between the metal and the n-type semiconductor when $\Phi_m < X$ and electrons are accumulated at the interface due to band bending. (b) Depletion. Electrons are depleted due to band bending when $\Phi_m > X$ and the Schottky barrier can be calculated by $\Phi_b = \Phi_m$ - X. (c) Fermi level pinning. The bands in the semiconductor bend before in contact due to surface states. The bands bend again after in touch with metal, followed by a contact barrier created at the surface. The barrier is caused by the metal induced gap states at the surface of the semiconductor, and the barrier height is independent of the metal work function. Φ_m, the metal work function; X, the electron affinity; Φ_b, the barrier height; E_C, the conduction band; E_F, the Fermi level; E_V, the valence band.

Thermal evaporation, also known as filament evaporation or resistive evaporation, is a process where metal is vaporized for deposition by resistive heating, a primary method of metallisation. Figure 3.11 shows a schematic of a typical configuration for a thermal evaporation chamber. When a material is to be

deposited, a solid source metal is placed in a ceramic crucible or in a basket/boat made of refractory material that has a very high melting point such as tungsten. Pumping of the chamber is required until certain vacuum pressure is reached (1×10^{-6} mbar), the source material is then heated by ramping up the current until a specified temperature is reached, dependent on the metal source material, where some vapour pressure is produced, followed by a short wait, as needed, to get rid of any possible contamination from the surface. A shutter is then opened for vaporized material to be able to traverse the chamber and reconstruct on the sample surface as a coating.

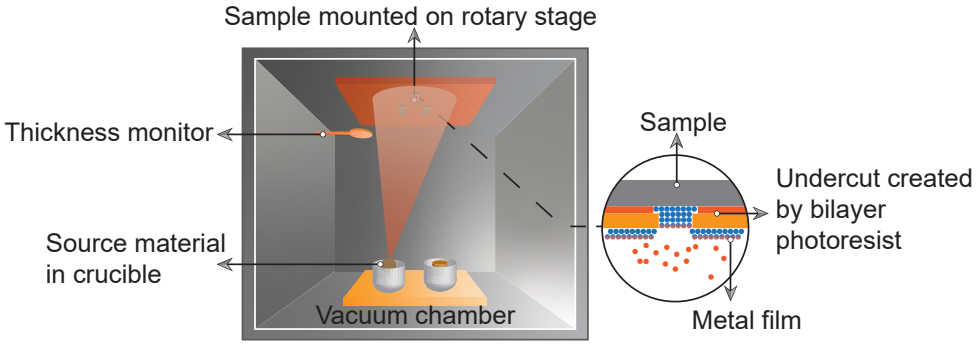

Figure 3.11: Diagram of a thermal evaporation process. The metal is vaporized by resistive heating and then reconstructs on the sample surface to accomplish film deposition.

Thermal evaporation offers high directionality, which may or may not be advantageous, depending on the device process required. A quartz crystal microbalance is in place to monitor the real-time thickness by frequency response during the coating process, which can be expressed by Sauerbrey's equation as

$$\Delta f = \frac{2{f_0}^2}{\sqrt{E\rho}} \times \frac{\Delta m}{A}, \tag{3.5}$$

where Δf is the frequency change, f_0 is the resonant frequency of the crystal, E is Young's modulus, ρ is the density, Δm is the deposited mass and A is the area that is piezoelectrically active. The shutter is then closed once the target thickness

is achieved, the current is gradually decreased, and the sample is ready when the chamber is fully vented for further operation. A lift-off process is typically carried out afterwards to remove metal on top of masked regions, to transfer the pattern from lithography onto the metal film. Bilayer photoresist is usually required to create the undercut profile for the lift-off process as shown in figure 3.11. An alternative approach for producing undercut structures is shadow evaporation that utilises angled deposition to achieve lateral or overlapping features.

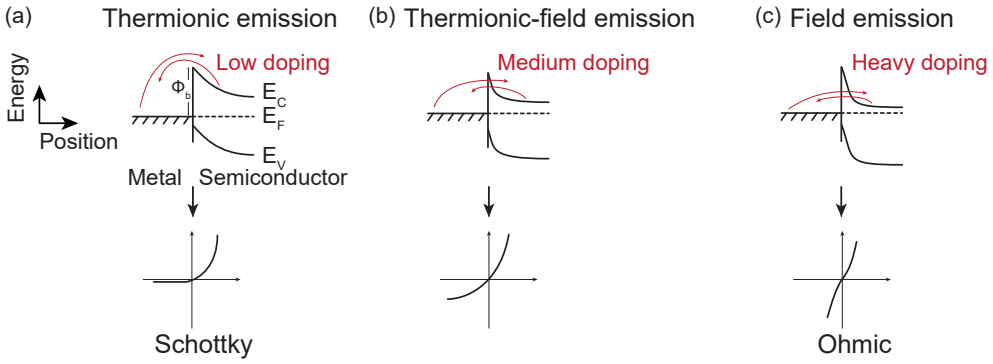

Figure 3.12: Three conduction mechanisms for metal-semiconductor interfaces of various barriers and corresponding I-V characteristics. (a) Thermionic emission. (b) Thermionic-field emission. (c) Field emission. Higher doping level shapes a more linear I-V curve. Φ_b, the barrier height; E_C, the conduction band; E_F, the Fermi level; E_V, the valence band.

Gold is accepted as a good contact metal for compound semiconductors due to its great conductivity and lack of oxide. Prior to gold deposition, an initial thin layer of Ti or Cr is widely used to improve substrate adhesion and metal-semiconductor interface as Ti (work function $\Phi \sim 4.3$ eV) or Cr (work function $\Phi \sim 4.5$ eV) can have a better work function matching with semiconductor. Figure 3.12(a)-(c) shows sketches of three conduction mechanisms as a function of the barrier height and width by the instance of metal/n-type semiconductor and corresponding I-V characteristics. Depending on the Fermi level of the metal and electron affinity energy of the semiconductor, the metal-semiconductor contact can be divided into

three types: Schottky, ohmic and a mixed one between the two. Thermionic emission is the major mechanism for an interface with Schottky barrier, while field emission can be explained for an ohmic interface which exhibits a more linear I-V curve. For the condition of a medium barrier interface, both thermionic and field emission contribute to the conduction. An ohmic contact is preferred and can be achieved by optimal metallisation and appropriate annealing.

In this work, a two-slot thermal evaporator from Moorfield minlab is used for Cr (rod), Ti (pellet in W basket) and Au (wire in alumina crucible) metallisation.

3.2.8 Sputtering

Metallisation of refractory material such as Nb with thermal evaporation can be difficult as they are highly resistant to heat, indicating extremely high currents and special crucibles are needed.

As an alternative, deposition via sputtering works well on these materials. In addition, sputtered films show more adhesion to the substrate and higher film density. With the assistance of a plasma, the sputterer delivers more isotropic deposition than thermal evaporation and a higher deposition rate for high throughput.

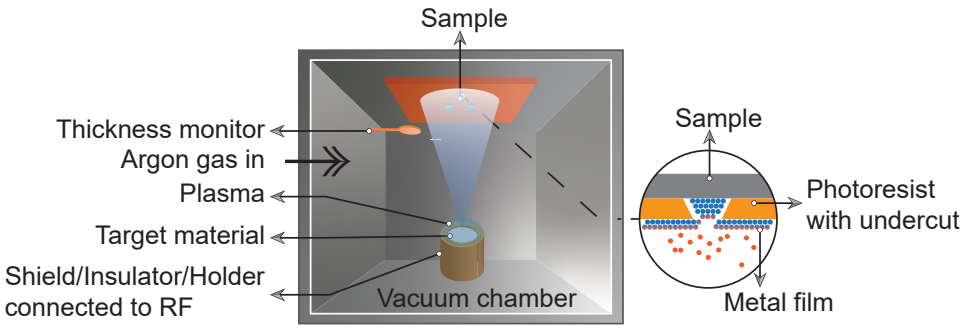

Figure 3.13: Illustration of a sputterer chamber. Sputtered metal atoms with high energy hit the sample surface to form a film.

Figure 3.13 shows a schematic of an RF sputterer. Inert gas such as argon is

pumped into chamber and when an RF is applied between the shield and the holder of the target, an argon plasma strikes near the top surface of the target, ionized argon molecules are attracted to the target and sputter metal atoms out. Those metal atoms then travel through the chamber with high velocity to reach sample surface top and form a thin film.

In this work, a two-slot sputterer from Moorfield with argon inert atmosphere is used for Nb and Ta film deposition.

3.2.9 Wire Bonding

Wire bonding is the process of bridging devices to peripheral circuits for measurement or electric functioning.

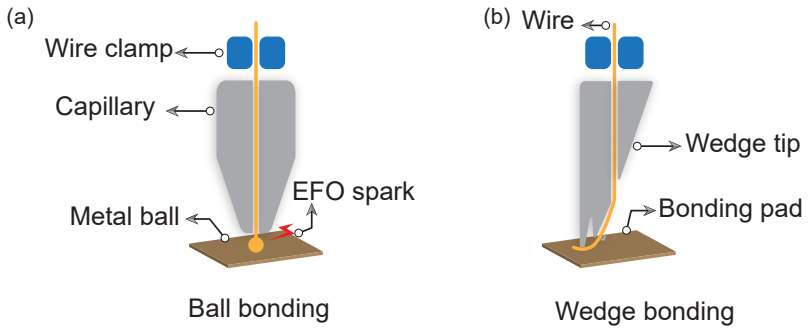

Figure 3.14: Schematic representation of two types of wire bonding, highlighting details of the difference between the tips. (a) Ball bonding. (b) Wedge bonding.

Figure 3.14 depicts two types of wiring: ball bonding and wedge bonding. In ball bonding, the wire is fed through a capillary. To start first bond, the 'electronic flame off' (EFO) melts the wire forming a ball at the tip of the capillary. The capillary is then placed into the bond position with ultrasonic energy and pressure applied until first bond is formed. The tip is lifted and move to second bond point while following an arched loop path, then the tip is lowered to be in touch with second bonding site. Afterwards, the wire clamp is closed and the capillary pulls back to break the wire to finish bonding, leaving a tail for the next ball formation. For

wedge bonding, firstly, the tip holding wire is positioned over the bonding site, once the tip is in touch with the site, ultrasonic energy is applied for a preset time until a firm connection is created. Secondly, the tip rises to loop height and moves in a loop forming a path, as with ball bonding, to the second bonding site. After second bond termination, the clamp is closed and the tip pulls back to break the wire while leaving a short tail for next bonding. In packaging that requires low profile loops, wedge bonding is preferred due to its straight line loop that is at a lower angle away from the surface, which can reduce the weakness in the heel of the bond while saving packaging space.

Ultrasonic power, force, time and bonding temperature are parameters to be considered for optimal bonding quality. Gold and aluminium are common choices as bonding wires. Ball bonding is most often used with gold wires while wedge bonding can do gold and aluminium. Wedge bonding is less susceptible to contamination than ball bonding but it may cause cratering underneath bond pad due to damage in those parts.

In this work, a TPT HB05 wire bonder with 25 μm thick gold wire is used for all wire bondings. Wedge bonding is used frequently, and ball bonding is used for fragile bonding pads. A SUSS RA 120M scriber is used for scribing before mounting chip to chip carrier for bonding.

3.3 Characterisation

3.3.1 Probe Station

A probe station is a platform where electrical measurements are carried out. As shown in figure 3.15, a typical configuration of a probe station contains an optical microscope for observation, a vacuum chuck for holding the sample in place, micro-manipulators where probes are mounted and measurement unit connected to probes.

Measurements on devices become challenging as the dimension shrinks. With micro-manipulators and fine-tipped probes, the probe station offers a high level of

accuracy for positioning and contact of small devices. Probes are made of tungsten (or copper alloy etc.) and have pointy tips to minimise the contact areas that are tiny bonding pads due to limited space on chips. Micro-manipulators with three-axis adjustment can offer higher spatial resolution for positioning the probes to the target contacts. The optical microscope provides visual assistance when locating sample and positioning probes and usually comes with a camera. A vacuum chuck is used to avoid movement of the chip during the test and provides better data repeatability. The electrical test is performed by a measurement unit connected to the probes, which can be source meter, etc.

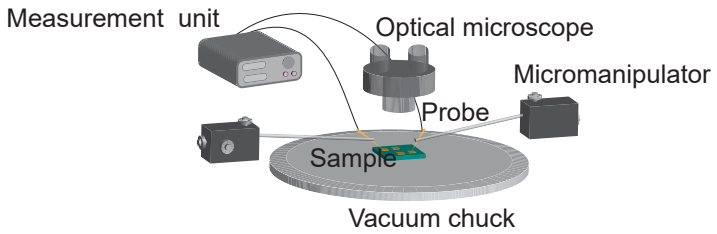

Figure 3.15: Schematic representation of components in a typical probe station set-up for electrical measurement.

A source measure unit (SMU) is an electronic instrument that sources and measures voltage and current at the same time. SMUs can be used in a range of tests from basic I-V sweeping to customized memory characterisation. The Keithley 2634B has two channels of voltage pulse sourcing, each of which samples the voltage and current simultaneously and also support four terminal measuring. With two channels present, both gate-source and drain-source can be measured while being pulsed in memory performance test.

One of the key parameters to be measured for the memory characterisation is the endurance cycles. Endurance is evaluated by repeatedly measuring the readout current following each programming and erasing cycle. The sequence of programming, readout, erasing, and subsequent readout is iteratively performed until the memory device exhibits failure. Endurance refers to the maximum

number of programming and erasing cycles a memory cell can reliably retain the memory window before it begins to degrade. This is a critical parameter for NVMs. High endurance ensures the longevity and reliability of memory devices. Endurance is affected by physical and electrical wear mechanisms, such as charge trapping, oxide breakdown, and material fatigue, which accumulate with repeated programming/erasing cycles. Retention test, on the other hand, denotes the ability of a memory cell to preserve stored data accurately over a given period without power supply.

The other key parameter is the retention time, which is an indispensable measurement of the non-volatility for memory devices. Retention is assessed by repeatedly sensing the channel current following a single programming (or erasing) operation, continued until the memory cell begins to exhibit signs of degradation. Retention performance is typically measured in terms of the duration for which a memory cell can maintain its programmed/erased state within specified error tolerances. For both endurance and retention test, the voltage pulses and current sensing for programming, erasing and readout are crucial for an accurate measurement of ULTRA**RAM**™. Other regular measurements, such as current-voltage characteristics, are also included to characterise the conductivity and other properties of the memory channel.

In this work, a probe station equipped with an SMU Keithley 2634B is used for electrical measurement. The 2634B SMU has a voltage programming resolution of 50 μV at 2 V range and a current measurement accuracy of 0.02% + 200 nA at 1 mA range. LabView software is used for measurement coding with SMU.

3.3.2 Surface Characterisation Technique

A scanning electron microscope (SEM) scans the sample using an accelerated and focused electron beam to provide imaging of a sample surface or morphological information. Electrons' interactions with the sample's atoms produces different signals containing information on sample topography and composition etc. Various

imaging and analysis techniques can be achieved with appropriate detectors. A common imaging technique is secondary electron imaging which gives information about topography.

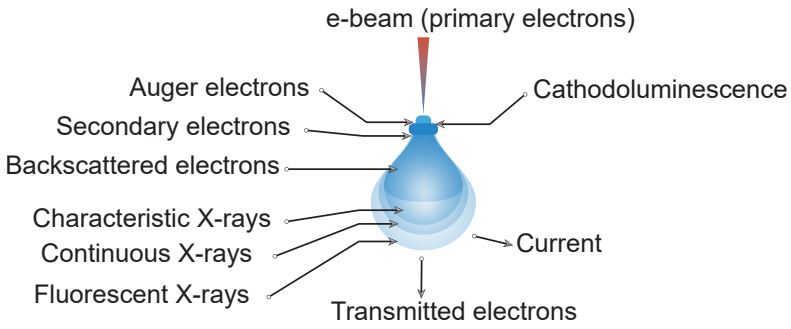

Figure 3.16: Schematic representation of the beam-specimen interaction, showing various signals that can be collected and analysed.

As an imaging tool, SEM operates on similar basic principles as an optical microscope, but it can achieve nanometre-scale resolution thanks to the short wavelength of electrons, providing extraordinary details of a sample without too much burden on specimen preparation. Figure 3.16 illustrates the signals generated by the electron beam hitting a specimen. It's worthy to note that resolution in scanning microscopy is usually determined by the interaction volume, which is typically larger than the beam size and also depends on the type of signal being detected. Another technique used in SEM is electron backscatter diffraction imaging, which gives the crystallographic structure of the specimen by detecting the backscattered electrons that are sensitive to orientation of grains in the sample. Moreover, the back-scattered electron signal can also be used for electron channelling contrast imaging (ECCI) which is a ubiquitous tool for defect imaging and analysis. Characteristic X-ray signals generated can also be used for chemical analysis and compositional characterisation, quantitatively or qualitatively, also known as energy-dispersive spectroscopy (EDS). X-ray photoelectron spectroscopy (XPS) is a surface-sensitive quantitative spectroscopic technique and provides information on the chemical composition of the top 10 nm (surface layer) of materials with $< 1\%$

precision. The XPS characterisation in this work was carried by Dr Samuel Jarvis and Alessio Quadrelli, both from the Department of Physics at Lancaster University.

A JEOL JSM-7800F SEM with EDS and a TESCAN SAFER (SEM Analysis with FIB, EDS and Raman) system are used for SEM imaging, EDS mapping and cross-sectional SEM imaging. An AXIS Supra X-ray photoemission spectrometer with radial distribution chamber, Ar ion gas cluster source is used for XPS characterisation.

3.3.3 Transmission Electron Microscopy and Focused Ion Beam

Transmission electron microscopy (TEM) is an analytical technique to reveal information in materials down to sub-nm scale. As TEM collects electrons transmitted through the sample, a specimen for TEM must be thin enough for sufficient electrons to pass. Typical sample preparation involves disc cutting, mechanical grinding, dimple grinding and ion milling. Specimens can also be prepared by focused ion beam (FIB), which employs a high-energy ion beam to customize the cutting and polishing in a specific region of the TEM sample.

A wide-range of imaging techniques are available for TEM, such as high-resolution TEM, scanning TEM and selected area electron diffraction. Modern aberration-corrected TEM can achieve atomic resolution. TEM imaging and measurement in this work were carried out by Prof. Richard Beanland and Francisco Alvarado Cesar, both from University of Warwick, and me.

3.3.4 X-Ray Nano-Probe Technique

X-ray nano-probe technique refers to a series of techniques using a hard X-ray as a nano-probe for analysis of nano-scaled structure or materials, including X-ray absorption near edge structure (XANES), X-ray fluorescence (XRF), X-ray diffraction, etc.

XANES derives from X-ray absorption spectroscopy. In a typical X-ray absorption spectroscopy curve, taking the absorption edge as a border, the entire spectrum can be divided into 3 parts: pre-edge part, near-edge part, which is XANES, and post-edge part. In general, XANES is about 10 eV below the absorption edge and 20 eV above the edge. XANES provides information on the local electronic structure of an atom as it evolves throughout a reaction. XRF is useful for determining the chemical composition in a variety of materials.

X-ray nano-probe analysis in this work is carried out at ID 16B in the European Synchrotron Radiation Facility.

3.4 Simulation

The nextnano software [237] allows a variety of simulations including electrical and optical properties across nanoelectronic and photonic devices. The software features a built-in Schrödinger-Poisson solver and a material database for III-V compounds with relevant physical properties, simulating band-structure, quantum well energies, wave-functions, electron/hole densities and current density etc.

The nextano non-equilibrium Green's function (NEGF) package is specifically designed for quantum transport simulation including density of states and current-voltage characteristics. nextnano multi-scattering Büttiker (MSB) method is based on a quantum transport method that follows the NEGF framework. It avoids self-consistent calculation of lesser self-energies by replacing them with a quasi-equilibrium expression. The nextnano++ package provides versatile computational capabilities, including calculation of band structures, densities of states and so on. The nextnano MSB package is included in the nextnano ++ package.

Simulations performed in this work are done with nextnano software using nextnano++ (with MSB method included).

3.5 Summary

To summarise, the techniques employed in the growth, fabrication, characterisation and simulation of ULTRA**RAM**™ are outlined and discussed. Notably, certain methods hold particular significance owing to their considerable impact on the performance and development of ULTRA**RAM**™. The growth quality of MBE forms the foundation for the fabrication of ULTRA**RAM**™ devices. The defect density, layer variation and uniformity across the grown wafer underpin the fabrication yield and the memory performance. With respect to the fabrication process, the most critical dimension of ULTRA**RAM**™ is the gate size, which is usually done in the first lithography step and patterned by a single exposure process. Therefore, the lithography technologies define the ultimate feature size achievable. On the other hand, ICP, particularly when accompanied by real-time monitoring, is the most critical etching technique, playing a pivotal role in determining the success of memory fabrication via building the high quality TBRT stack and the channel thinning with nanometre precision. In the context of characterisation, TEM provides essential insights, serving as both validation and quality assessment of the wafer growth and fabrication processes of ULTRA**RAM**™. TEM imaging offers valuable feedback for the iterative optimisation of the ULTRA**RAM**™ fabrication process and device design.

Chapter 4

Fabrication and Scaling of ULTRARAM™

4.1 Scaling Scheme

The ultimate objective of the scaling is 50 nm while the current state of the art of ULTRA**RAM**™ demonstrated was 20 μm in terms of the feature size, i.e. the gate dimension. To make the way to the nanometre scaling of ULTRA**RAM**™, the targeted scaling scheme is divided into three steps in terms of feature size, as shown in figure 4.1. Micron, sub-micron and nanometre, with the employment of optical mask aligner, LW and e-beam lithography (EBL), respectively. The scaling route is based on the balance of capability, time-efficiency, cost and availability of each lithography approach.

Previously demonstrated prototype devices were achieved with gate sizes ranging from 20 μm to 30 μm, and were fabricated by a standard optical mask lithography. Therefore, to start with, a mask aligner was used as an initial attempt for scaling with processing changes only while device geometry and fabrication flow remain unchanged. In the next stage, an LW is required as device re-design is inevitable for a scalable structure, and an EBL is necessary for further scaling where the capability for fine features and alignments is imperative. The LW offers flexibility of patterning

with mask-less capability, but the minimum feature is limited to around 500 nm, serving as an appropriate apparatus for the transition stage between mask aligner and EBL in regard to the feature size.

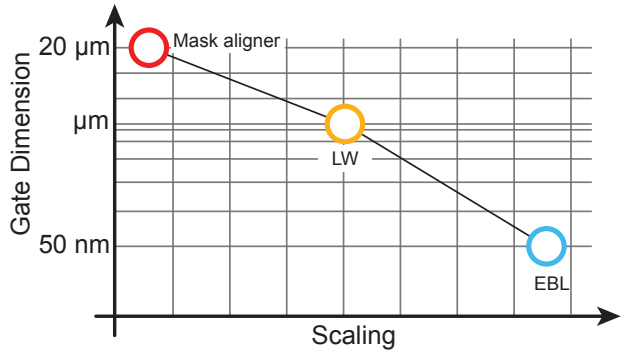

Figure 4.1: Sketch of the scaling route for ULTRA**RAM**™ memory.

The aim of the first stage of scaling is a scalable (dry) processing method. As the previously used method of wet etching is not suitable for scaled devices when the gate dimension approaches below micron: the lateral etching from wet-etch will be comparable to vertical etching and becomes detrimental to device fabrication. The second stage focuses on the design of the device to increase the feasibility of the downscaling plan, and the improvement of device performance by overcoming the downsides from the legacy fabrication method.

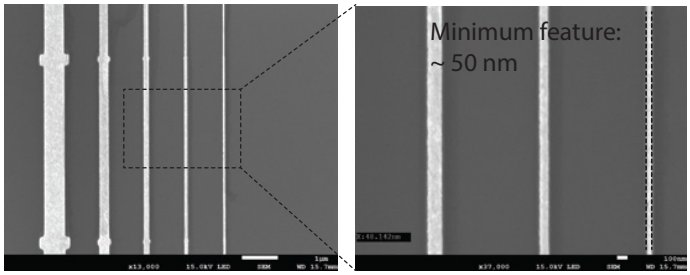

Figure 4.2: SEM images of a group of EBL patterned metal bars, showing the achieved minimum feature size $\sim$ 50 nm.

With all design and processing optimised and verified, the device can be eventu-

ally demonstrated with practicable minimum feature size and decent performance using EBL. The EBL has been validated with a patterning test and demonstrated the capability of lithography for below-100 nm feature sizes. As shown in figure 4.2, a group of metallisation bars patterned by EBL shows a minimum width $\sim$ 50 nm.

4.2 Fabrication and Self-Aligned Design

Details of the evolution of ULTRA**RAM**™ memory fabrication are provided here. The earlier version using wet-etching is discussed first, followed by the scalable version featuring a more self-aligned design and all-dry etching. Finally, the latest structure design is described with an improved version of the self-aligned process, featuring misalignment tolerance, as well as compact geometry design for enhanced memory performance.

4.2.1 Process Flow with Wet-Etching

Wafer Layout

The growth structure of the memory wafer used for the processing with wet-etching is outlined in table 4.1. The III-V layers were grown by Dr Peter Hodgson by epitaxy on a Si substrate. The memory layers ranging from floating gate to back gate are grown on top of the wafer, isolated from the substrate by buffer layers for tackling lattice mismatch and filtering dislocations. The initial AlSb layer on the Si substrate forms into 3D islands that reduce the diffusion length of Ga atoms during the initial growth of the sub- sequent GaSb, facilitating the nucleation of the overlying GaSb buffer into a 2D layer and preventing the formation of planar twinning defects. To expand on this point, the 3D islands localize the misfit strain, thereby reducing the density of threading dislocations propagating upward. In particular, AlSb 3D islands create many independent nucleation centres, which lead to a more uniform film in the subsequent growth process. Moreover, the 3D AlSb islands can break the long-range registry, thereby helping to suppress the formation

of antiphase boundaries.There are two major reasons for using GaSb as a buffer layer rather than InAs. Firstly, GaSb has shown antiphase boundary-free growth on Si, which helps to reduce the antiphase boundary when growing polar InAs/AlSb layers onto the non-polar Si substrate. Secondly, GaSb has the least mismatch to both InAs and AlSb, accommodating the large Si mismatch while providing a nearly lattice-matched platform for the subsequent InAs/AlSb growth. The channel material is n-doped InAs which indicates the normally-on nature of ULTRA**RAM**™ memory.

Layer	Material	Nominal thickness (nm)	Thickness measured by TEM (nm)
Floating gate	InAs	10	9.7
Quantum barrier	AlSb	1.8	2.5
Quantum well	InAs	2.4	1.8
Quantum barrier	AlSb	1.2	1.8
Quantum well	InAs	3.0	2.5
Quantum barrier	AlSb	1.8	2.5
Channel	InAs (n-type 5×10^{18})	10	10.7
	GaSb	20	19.6
	AlSb	8	7.2
Back gate	InAs (n-type)	50	58.7
2-step buffer	GaSb (hot)	540	-
	GaSb (cold)	1400	-
Nucleation	AlSb	5.2	-
Substrate	Si	$\sim$ 380 μm	-

Table 4.1: Detailed layout of XPH 1452 wafer for ULTRA**RAM**™, utilised in the wet-etching fabrication.

TEM characterisation of the wafer is shown in figure 4.3. The buffer layers block the vast majority of defects from the lattice mismatch at bottom interfaces, with a minority of threading dislocations sprouting into the top memory layers. The unique TBRT structure grown in high quality is observed. The measured thicknesses for each layer using TEM are collected in table 4.1. In comparison to the target design, the actual thickness shows minor deviation from the original design. The dashed boxes in figure 4.3 indicate regions for zoom-in TEMs. Colourful dots placed in various layers represent the relevant III-V layers denoted in the legend. The observed

TBRT structure with sharp InAs/AlSb interfaces suggests a successful MBE growth of the XPH 1452 wafer.

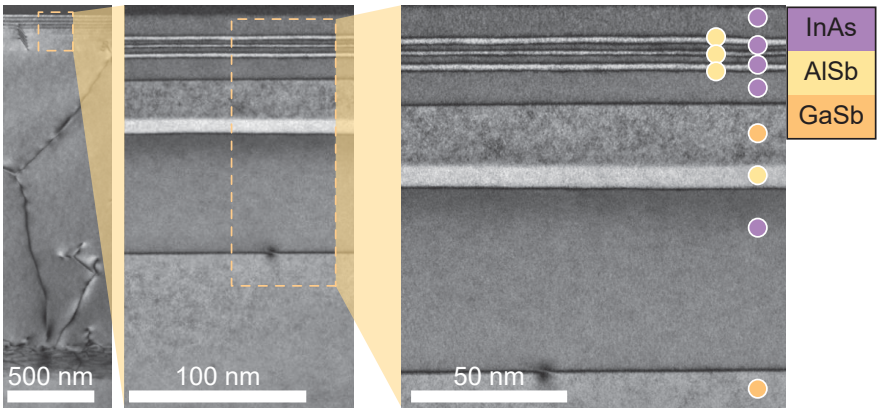

Figure 4.3: Cross-sectional TEM images of different scales of XPH 1452 wafer, presenting high quality of III-V memory layers. The coloured dots denote the corresponding materials in the legend. Images provided with permission from Dr Richard Beanland, University of Warwick.

Fabrication Steps

The entire fabrication flow is sketched in figure 4.4(a)-(i) and corresponding experimental photos of a 30-μm gate device for each step in processing are shown in figure 4.5(a)-(h).

The fabrication starts with a pre-cleaning of the cleaved wafer using ultrasonic agitation in acetone and isopropyl alcohol (IPA), followed by a baking at 110 °C for 5 minutes to remove moisture before any processing. The initial step, as illustrated in figure 4.4(a)-(b), is to construct a mesa to isolate individual devices. The sample is spin-coated with S1813 photoresist, followed by an edge beading-removal exposure and developing wherever needed for a better contact to the optical mask in subsequent exposures. Next comes the mesa exposure using the mask aligner and the pattern is transferred to the photoresist after a developing operation. In the meantime, alignment marks intentionally designed on the mask are printed onto photoresist as well, for alignments used in the following exposures. Then, the sample

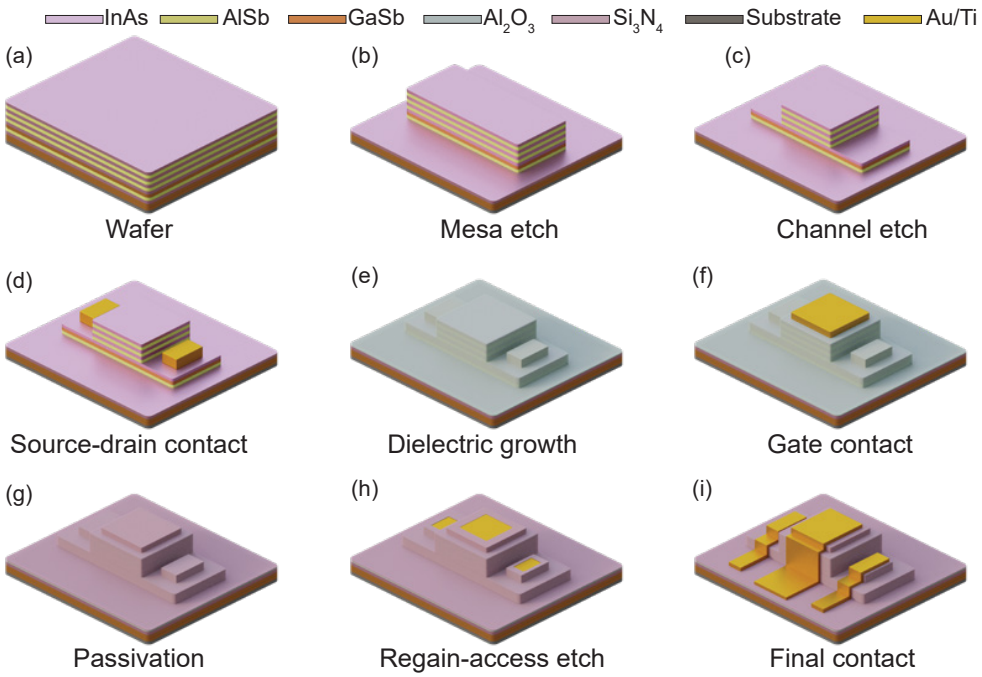

Figure 4.4: Simplified illustration of fabrication steps for wet-etching design that requires five alignments. (a) Wafer preparation. (b) Mesa etch. (c) Channel etch. (d) Source-drain contact formation. (e) Dielectric growth. (f) Gate contact deposition. (g) Passivation layer deposition. (h) Regain access etch. (i) Final contact formation.

is loaded into the ICP chamber to be etched by mixed gaseous etchants, including Cl_2, BCl_3 and Ar. Chlorine-based recipes [238] are widely used for III-V dry etching and the addition of BCl_3 is useful for etching the native oxide on the top surface. The choice of ICP is by the reason of the fast etching rate provided by the higher plasma density. In terms of etching control and monitoring, a laser emitted from the end-point detection kit is directed at the region to be etched. The focus and aperture are adjusted to accommodate any reflectance intensity changes during the etching process. The etching is stopped in the target layer, as guided by the simulation result. In this step for mesa formation, it's the InAs back gate layer that the etch must be stopped in. After photoresist stripping and appropriate ashing to remove any existing residue, the mesa is formed on the sample as shown in figure

4.4(b). Prior calibration, typically comprising the collecting the etching curve of the etching-through of a sample with confirmed layer structure and the comparison to the simulated reflectance, is required before any subsequent processing may take place. By following the characteristic signs on the etching curve, the accurate etching of multiple layers is achievable. The typical etching rate is around 23 nm/min for channel etching and 71 nm/min for mesa etching.

The channel etching in figure 4.4(c) is the most crucial part of the entire fabrication due to the atomically-thin structure where an accurate and well-controlled operation is required. Similar to the lithography process aforementioned, a source-drain pattern layer on the mask is used for channel patterning. A bilayer S1813/LOR 3A is spin-coated to create an under-cut structure for lift-off purposes. An alignment is needed in this step to ensure the exposure window is positioned correctly within the mesa region. This is done by a typical two-step alignment with the mask aligner, where large and fine alignment marks printed in the previous step are used in succession. To reach the channel layer, wet etching is used here. The sample is put into citric acid solution, deionised (DI) water, MF-319 developer (tetramethylammonium hydroxide-based), DI water sequentially to selectively etch the InAs and AlSb, respectively. Then this is repeated to remove all layers in the TBRT until the InAs channel layer is revealed. The photoresist is removed later by remover 1165 and the source-drain steps are shaped on the defined mesa, as shown in figure 4.4(c).

The source-drain contact is patterned and metallised via thermal evaporation. A lithography with source-drain contact pattern is carried out first where a second alignment operation is needed. A gentle ashing prior to evaporation helps to clean the processed surface and to provide a better adhesion for subsequent metal depositions. Thereafter, a thin Ti film of 10 nm is first deposited to enhance the adhesion of the metal to the substrate and to deliver ohmic contact. As shown in figure 4.6, the work function of Ti is 4.3 eV, and is smaller than the affinity of InAs which is 4.9 eV. This forms an ohmic contact at the interface, while the direct

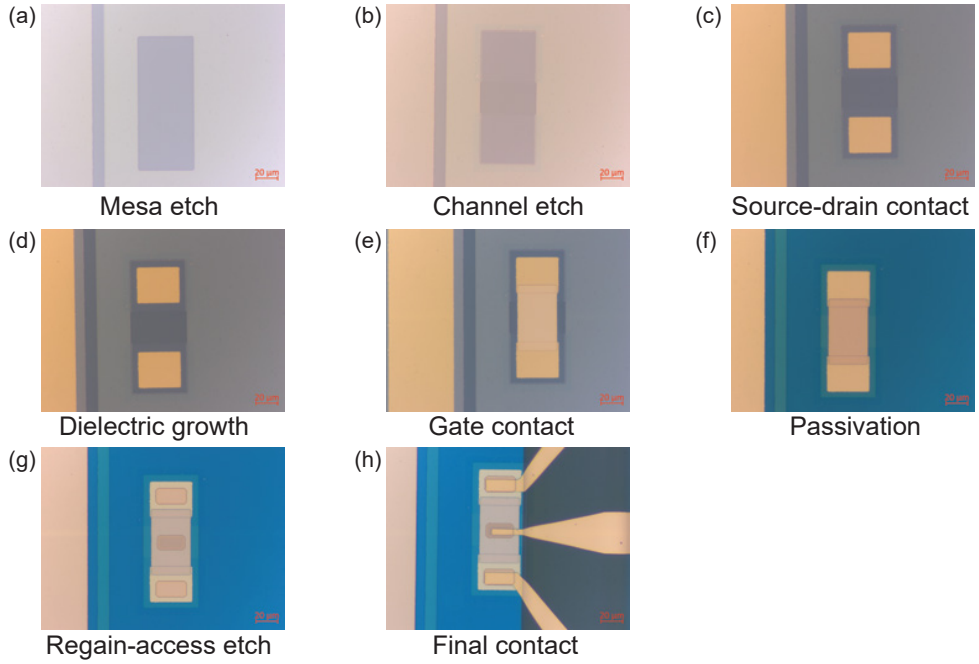

Figure 4.5: Optical photos from each step in the fabrication using the wet-etch design. (a) Mesa etch. (b) Channel etch. (c) Source-drain contact formation. (d) Dielectric growth. (e) Gate contact deposition. (f) Passivation layer deposition. (g) Regain access etch. (h) Final contact formation.

contact of gold ($\Phi \sim 5.3$ eV) to MBE grown InAs (100) has a barrier height of 0.5 eV [239]. Literature shows that Fermi-level pinning for InAs is about 0.13 eV above conduction band [240, 241]. Then, a gold layer around 50 nm is deposited to provide the optimal conductivity, forming the source-drain contact shown in figure 4.4(d) after a lift-off process which removes all the metal film sitting on photoresist-covered region by assistance of an under-cut structure. The evaporation of the two metals is carried out consecutively under high vacuum conditions. To avoid the oxidisation of the InAs channel caused by exposure to air, the sample is stored in N_2 while not in processing.

Ashing is required prior to dielectric growth in the ALD chamber to remove any contamination from sample processing. As shown in figure 4.4(e), the dielectric layer is completed by a conformal growth of alumina via ALD which also covers

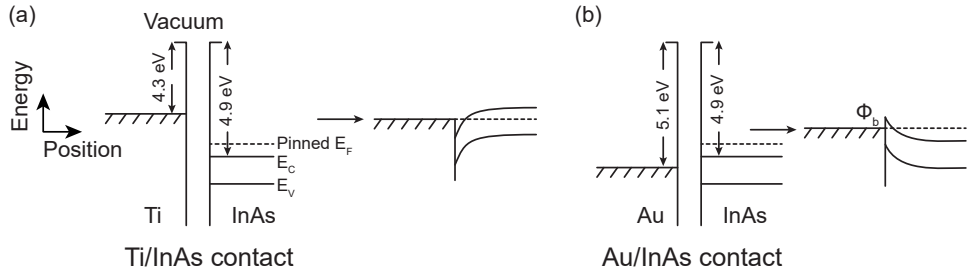

Figure 4.6: Simplified band diagram of (a) Ti and (b) Au contact to InAs channel. A Schottky barrier occurs for the use of gold. E_F, the Fermi level; E_C, the conduction band; E_V, the valence band; Φ_b, the barrier height.

the sidewall of the device to provide protection. The choice of ALD alumina is attributable to the self-limiting property which allows precise control of film thickness, as well as the film quality. Alumina is selected as dielectric by its high electrical insulation and permittivity. A 15-nm Al_2O_3 layer grown by 150 cycles of water-TMA pulse sequence at 150 °C produces amorphous alumina with a high breakdown field strength of 8.3 MV/cm [242], which is of vital importance for dielectric purposes. A temperature of 150 °C is also preferable for the channel, as higher temperatures may lead to As desorption from the InAs channel, resulting in an In-rich layer that adversely affects channel conductivity. The growth rate is around 1 Å per cycle under such conditions. The 15-nm thickness is determined by the requirement of memory operation as a thicker alumina layer will incur higher program/erase voltage while a thinner dielectric layer can break down easily at a lower voltage and may exhibit leakage.

At this point, the sample is ready for gate contact fabrication. Standard optical mask lithography with a third alignment transfers the gate contact pattern from the mask onto the bilayer photoresist. The same metallisation as in source-drain contact is carried out to deposit the gate metal as illustrated in figure 4.4(f). The following step in figure 4.4(g) deposits Si_3N_4 to provide a final passivation to the device. Employment of PECVD here rather than ALD is based on the deposition

rate and requirement of thicker protection layer. PECVD can achieve a deposition rate of ~ 60 nm/min for Si_3N_4. A three-minute deposition produces a 180 nm Si_3N_4 film which is enough for general passivation.

Figure 4.4(h) depicts the access window opened on Si_3N_4 to bridge the final metallisation with the source-drain contact and the gate contact completed in previous steps. This is done by the mask aligner using an etching window pattern with the fourth alignment, followed by RIE etching. Two layers of S1813 are used to counter the long-time aggressive etching. A two-step RIE etching is performed in this process. The material removal at the source-drain contact region consists of Si_3N_4 and Al_2O_3, while only Si_3N_4 layer at the gate region needs to be etched. Therefore, Si_3N_4 is etched first with a CHF_3 recipe, then Al_2O_3 is removed by a CF_4 recipe. Since the etching recipe is relatively selective over gold, the gate metal previously deposited can stop the over-etching at the gate window region. A 30-second duration is added up to etching time calculated from the estimated etching rate to ensure the full removal of Si_3N_4 and Al_2O_3 because the RIE etching cannot be monitored in real-time.

The final contact in figure 4.4(i) marks the completion of the entire fabrication process. A final optical mask lithography with the fifth alignment delivers the patterning and is followed by an evaporation of 20 nm of Ti and 160 nm of gold. To ease the lift-off process, the LOR 3A photoresist is recommended to be thicker than the metal film by around 25%. The thickness of LOR 3A used in this work is around 300 nm, and for bonding reasons, a thicker gold layer is necessary. A contact-lifting layer using hard-baked S1813 can be used to mitigate the height difference of the gate stack and mesa floor for a better contact connection. Also, a further metallisation at the contact pad region can be helpful for wire bonding. However, these two steps will introduce another two alignments, increasing the complexity of fabrication flow and reducing the quality due to accumulative alignment errors.

4.2.2 Scalable Route with All-Dry-Etching

The isotropic wet-etching causes the lateral removal of material, as illustrated in figure 4.7. This works well with large size devices thanks to the small height/width ratio, but it gets worse for scaled devices where the width is significantly reduced, and the detrimental lateral etching cannot be neglected. Therefore, dry-etching is introduced to all etching steps to develop a scalable fabrication route for ULTRA**RAM**™ fabrication. Depending on the dry etching conditions, the resulting sidewall profile may be either sloped or vertical. Typically, an increase in gas flow, chamber pressure, and coil power tends to produce a sloped etch profile. In contrast, a more vertical profile can generally be achieved by increasing the platen power.

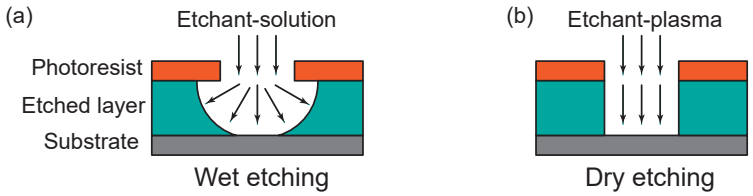

Figure 4.7: Comparison between (a) wet etching and (b) dry etching. Dry etching is more directional.

In the all-dry etching flow, the ICP etching with end-point detection enables the real-time monitoring for precise control of the etching process. In addition, there are a total of five alignments in the wet-etching assembly flow, and the fabrication yield is restricted by cumulative alignment mismatch. Hence, a (partly) self-aligned fabrication with fewer alignments is desired. A hard mask of Si_3N_4 is used in this fabrication route, reducing the number of total alignments to three, simplifying the assembly footprint. Consequently, the entire working flow is re-designed and optimised to facilitate scalable manufacturing.

Epitaxial Design

The epitaxial design used for the self-aligned device structure is outlined in table 4.2, which includes the target thickness and the measured thickness from TEM measurements, showing minor deviation from the original design. As an

improvement, XPH 1823 incorporates four repeating isolation units between memory layers and buffer layers that contain AlSb/GaSb/$Al_{0.7}Ga_{0.3}Sb$/GaSb as a barrier to block the vertical p-type leakage in the device. AlSb has a higher bandgap to block the leakage, but due to its susceptibility to degradation upon exposure to air, an additional layer of $Al_{0.7}Ga_{0.3}Sb$ which adds Ga in to mitigate the oxidation process is used as the actual mesa stopping layer. The GaSb layer inserted in between is to prevent relaxation as AlSb and $Al_{0.7}Ga_{0.3}Sb$ are both compressively strained on GaSb thus contribute towards reaching the critical thickness. Multiple wafers are used in this stage with minor changes to XPH 1823 while the majority of the design remains unchanged.

Layer	Material	Nominal thickness (nm)	Thickness measured by TEM (nm)
Floating gate	InAs	10	8
Quantum barrier	AlSb	1.8	2.2
Quantum well	InAs	2.4	3.3
Quantum barrier	AlSb	1.2	1.6
Quantum well	InAs	3.0	2.
Quantum barrier	AlSb	1.8	2.2
Channel	InAs (n-type 5×10^{18})	10	9.2
	GaSb	20	-
Isolation × 4	AlSb	8	-
	GaSb	50	-
	$Al_{0.7}Ga_{0.3}Sb$	30	-
	GaSb	50	-
2-step buffer	GaSb (hot)	540	-
	GaSb (cold)	1400	-
Nucleation	AlSb	5.2	-
Substrate	Si	$\sim$ 380 μm	-

Table 4.2: Detailed layout of XPH 1823 wafer for ULTRA**RAM**™, featuring the introduction of isolation layers.

The TEM characterisation of XPH 1823 wafer at different scales is shown in figure 4.8(a), depicting high quality memory layers including the TBRT and isolation, but still some dislocations can be found in bottom layers as shown in the large scale TEM image on the left. The coloured dots denote a clear structure of the TBRT.

The ECCI image in figure 4.8(b) presents the surface condition of the as-grown wafer, showing loosely distributed stacking faults and dislocations in a large area.

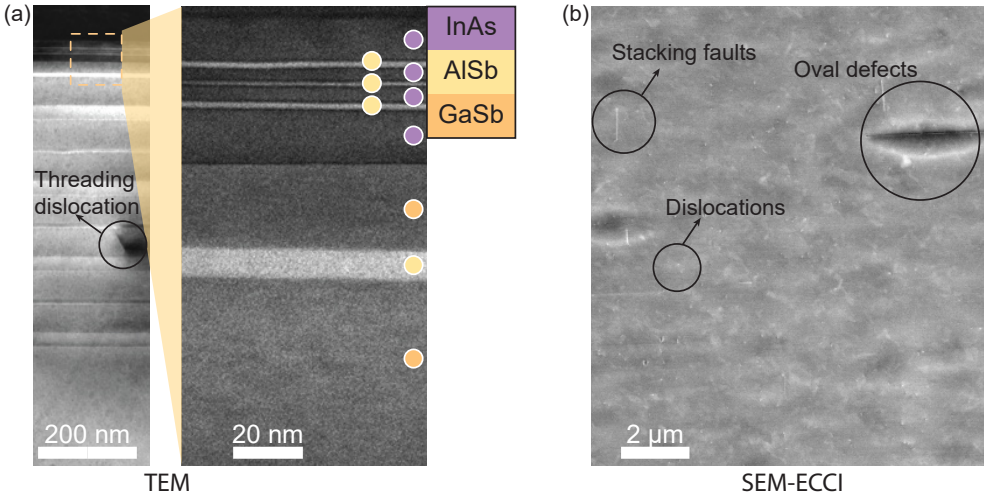

Figure 4.8: (a) Cross-sectional TEM images at different scales and (b) SEM-ECCI image of XPH 1823 wafer, presenting the vertical structures and surface condition of as-grown wafer, respectively. The coloured dots denote the corresponding materials in the legend. Images provided with permission from Dr Richard Beanland, University of Warwick.

Fabrication Steps

The self-aligned processing is illustrated in figure 4.9(a)-(i) and the corresponding photos for each step of a 30-μm gate size device are shown in figure 4.10(a)-(h).

As represented in figure 4.9(a), after a pre-cleaning process, alumina, Ta and Si_3N_4 are deposited to construct the first hard-mask layer by ALD, sputtering and PECVD, respectively. Dielectric alumina is deposited by 150 cycles of ALD growth, followed by 80 nm of Ta, and finally covered by 60 nm of Si_3N_4. Ta is selected as the gate metal due to its high stiffness and etch resistance to work as a hard mask, the top Si_3N_4 is to protect the Ta sputtering in the following steps to avoid leakage. Sputtered Nb, Ta, thermally evaporated Cr and, eventually, Si_3N_4 are used in different batches of fabrication to address the gate leakage issue, which will be discussed in section 5.1 in detail. As depicted in figure 4.10(d), to tackle

the inevitable misalignment, the mesa pattern is designed to straddle over the gate stack, thus causing two uncovered regions on both ends of the gate stack exposed to aggressive mesa etching in the next step.

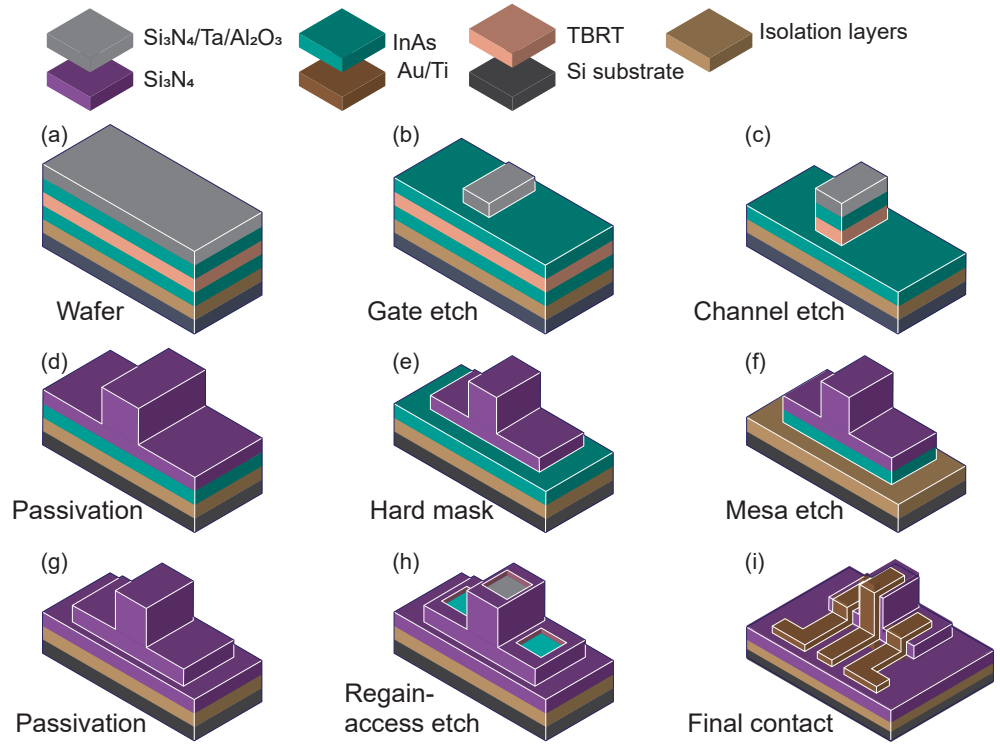

Figure 4.9: Schematic representation of the self-aligned design with fabrication steps. (a) Hard mask layers deposition on wafer. (b) Gate definition etch. (c) Self-aligned channel access etch. (d) Passivation with Si_3N_4. (e) Hard mask definition etch. (f) Self-aligned mesa etch. (g) Second passivation layer. (h) Regain access etch. (i) Final contact formation.

Figure 4.9(b) defines the gate stack by LW lithography with double layers S1813 and a consecutive etching that consists of Si_3N_4, Ta and alumina etching, forming the hard mask for the following channel etching. This is done in an RIE chamber using three different recipes, preparing the sample ready for the ICP channel revealing. A descum process by oxygen plasma is carried out before the etching to improve the sidewall profile. All RIE etchings here are non-chlorine based which are slightly

selective over the InAs channel. All etchings are done with an additional 30 seconds added into the calculated etching time based on etching rate to ensure complete removal of the three layers.

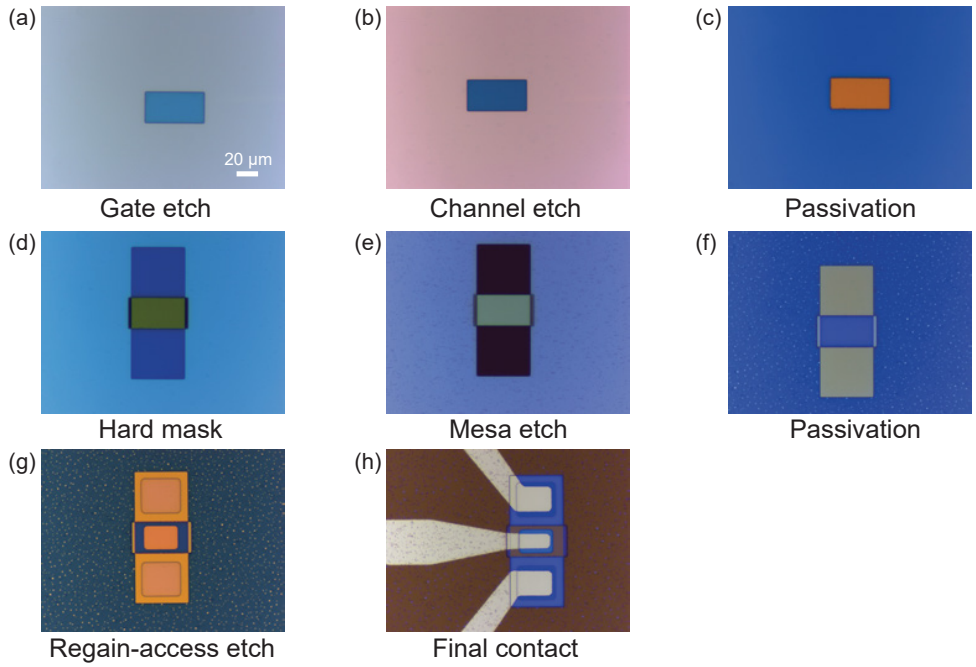

Figure 4.10: Fabrication photos of each step in the self-aligned design. (a) Gate definition etch. (b) Self-aligned channel access etch. (c) Passivation with Si_3N_4. (d) Hard mask definition etch. (e) Self-aligned mesa etch. (f) Second passivation layer. (g) Regain access etch. (h) Final contact formation.

Figure 4.9(c) depicts the first self-aligned process of the fabrication. The photoresist is removed beforehand to avoid the hardened S1813, which is hard to remove afterwards, and to reveal the previously deposited Si_3N_4 layer as the hard mask. Then, the channel layer is revealed by ICP etching with the assistance of the defined hard mask. Also, the dry-etching here enables the channel etching for even smaller features by eliminating the deleterious lateral etching from wet etching method, marking the establishment of a scalable route. Since the channel contains only 10 nm of InAs, a well-controlled etching with end-point detection technique using laser interferometry is applied here. Both CH_4-based and BCl_3-

based recipes are used in different fabrications to improve the etching control, more details are provided in the following section 4.3 on channel etching and end-point detection. Figure 4.11(a) presents the reference point to stop the channel etching for a fabrication with XPH 2213 wafer, exhibiting a high degree of correspondence with the simulation in figure 4.11(c). A photoresist stripping process that involves acetone, IPA and ashing is performed afterwards to finish up the channel etching.

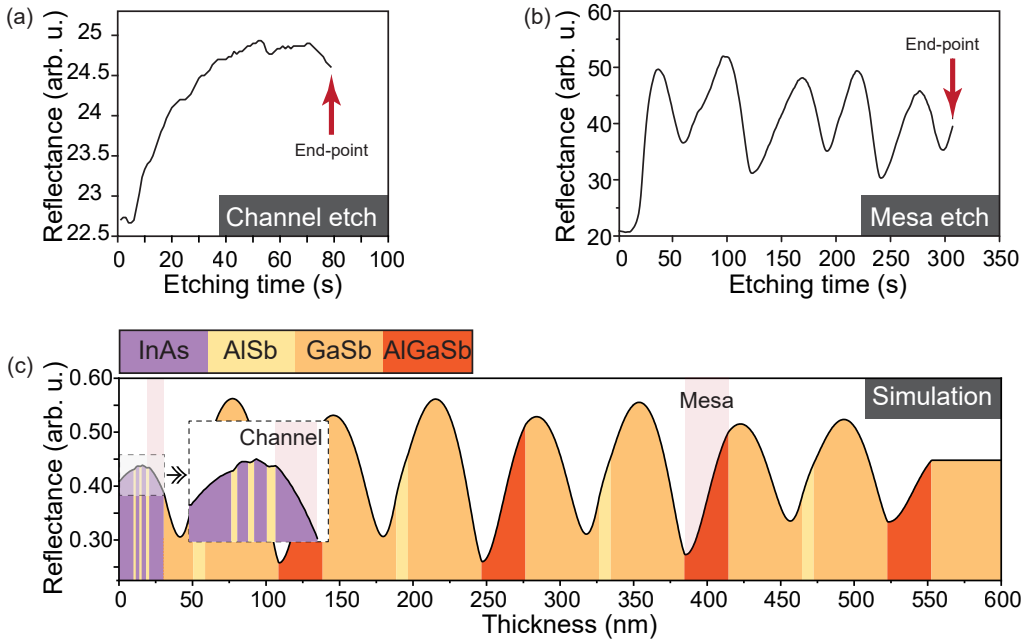

Figure 4.11: Etching reflectances from the fabrication for (a) channel etching and (b) mesa etching, as a comparison to reference regions indicated in (c) simulated curves. The inset in (c) shows the zoom-in of the TBRT and channel region.

A thick layer of $\sim$ 180 nm of Si_3N_4 is deposited in figure 4.9(d) to protect the exposed TBRT sidewall, also serving as a second hard mask. Optical mask lithography is used to pattern the mesa with single layer of S1813. The Si_3N_4 hard mask is constructed after RIE etching and photoresist stripping as depicted in 4.9(e). The second self-aligned etching comes in figure 4.9(f) where the mesa is formed by an ICP etching until the third repeating $Al_{0.7}Ga_{0.3}Sb$ layer. The end-point detection result is shown in figure 4.11(b), as a strong alignment with its simulation in figure

4.11(c). A polymer removing process is required for the fabrication using CH_4-based etching recipes as the polymer builds up over time and blocks further removal of mesa layers.

Final passivation is done with another PECVD Si_3N_4 deposition with thickness of $\sim$ 180 nm, covering everywhere of the device to protect it against ambient environment in future tests, as depicted in figure 4.9(g). The following step opens windows for final contact to the device by standard optical mask lithography with single layer of S1813 and RIE etching. The etching time here needs to be carefully calculated. The Si_3N_4 removal at source-drain region consists of two layers of Si_3N_4, while three layers of Si_3N_4 were deposited at the gate region, therefore, the general etching time is determined by the thicker one. Since the etching recipe is relatively selective over InAs, the access windows can be achieved with appropriate etching time as illustrated in figure 4.9(h). As a practical tip here, usually another thin layer of 60 nm Si_3N_4 is deposited right after opening the windows to protect the revealed channel from attacking by the developer in the next step which would involve an extra Si_3N_4 etching before final contact. As shown in figure 4.9(i), the device is finished after the final lithography with alignment followed by a thermal evaporation of 20 nm of Ti and 80 nm of gold. The fabrication ends upon the finishing of the lift-off procedure. As shown in figure 4.10(f), bubble-like dots appear after the second passivation step, this is likely to be caused by the polymer residue from mesa etching. Since this happens on the mesa floor, it does not affect the memory device directly. However, it has a negative effect on the wire bonding which becomes fragile due to the poor adhesion on the surface, a factor contributing to the eventual dropping of CH_4-based recipes and the switching to BCl_3-based ones.

4.2.3 Improved Self-Aligned Structure

Critical problems including high parasitic resistance and limited electrostatic gate control remain to be overcome in the self-aligned design. This will be discussed in detail in section 5.3.2. Figure 4.12 shows a coloured SEM of a device fabricated

by the self-aligned design, showing a gate-stack-source/drain gap in the device structure, this, in combination with the inevitable over-etching from the device processing, led to the poor channel readout performance which consequently causes the small memory window.

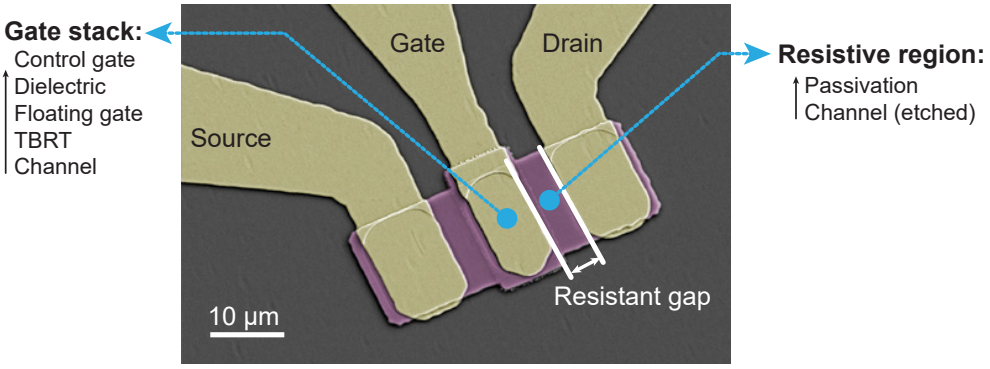

Figure 4.12: Coloured SEM image of a device fabricated with self-aligned design, a gap is shown between the gate stack and the source/drain contact. The resistant gap of micron scale originates from the device design. Layer arrangements for the gate stack and the resistive region are listed on both sides of the image. The gold colour denotes the region covered by contact metal and the magenta region represents the surface covered by passivation Si_3N_4.

Thus, a fully new device structure and a corresponding processing flow are proposed for fabrication, featuring a self-aligned design and a compact geometry while maintaining merits from legacy method. Moreover, the new design is misalignment-tolerant, eliminating the requirement for a high-precision alignment in the lithography that is a prerequisite for earlier versions of the device processing. Furthermore, this is the first fully mask-less fabrication design, empowering flexible designs on demand. Four alignments in total are needed in this design. A minimum gate dimension of 5×5 μm^2 is included in the design.

Growth Structure

As represented in table 4.3, the XPH 2318 wafer used for the improved self-aligned design employs a thicker InAs channel layer of 15 nm as an improvement

for better etching control, and an undoped channel offers smaller channel resistance. The increased thickness of the InAs channel provides greater tolerance to the over-etching issue and results in higher conductivity due to reduced surface scattering compared to the 10 nm InAs layer. Layer thicknesses obtained by the cross-sectional TEM measurement on an unprocessed wafer are also included here to show the difference from the target thickness. InAs layers in the TBRT are slightly thinner than the design value while AlSb layers are thicker than the nominal thickness. The floating gate layer is increased to 15 nm in an effort to enhance the charge storage capability for a greater memory window. The TBRT layers remain the same as the previous design.

Layer	Material	Nominal thickness (nm)	Thickness measured by TEM (nm)
Floating gate	InAs	15	13.061
Quantum barrier	AlSb	1.8	2.23
Quantum well	InAs	2.4	2.336
Quantum barrier	AlSb	1.2	1.805
Quantum well	InAs	3.0	2.442
Quantum barrier	AlSb	1.8	2.23
Channel	InAs	15	15.398
	GaSb	20	19.991
Isolation × 4	AlSb	8	8.178
	GaSb	50	-
	$Al_{0.7}Ga_{0.3}Sb$	30	-
	GaSb	50	-
2-step buffer	GaSb (hot)	540	-
	GaSb (cold)	1400	-
Nucleation	AlSb	5.2	-
Substrate	Si	$\sim$ 380 μm	-

Table 4.3: Detailed layout of XPH 2318 wafer for ULTRA**RAM**™, incorporating undoped InAs channel.

High resolution TEM characterisations of the wafer are shown in figure 4.13. Defects such as dislocations are observed at the bottom layers in the overview image of the entire vertical structure, but are effectively suppressed by buffer layers and isolation units as much less defects were seen on the top memory layers. The close-up image on the right shows high quality TBRT layers as annotated.

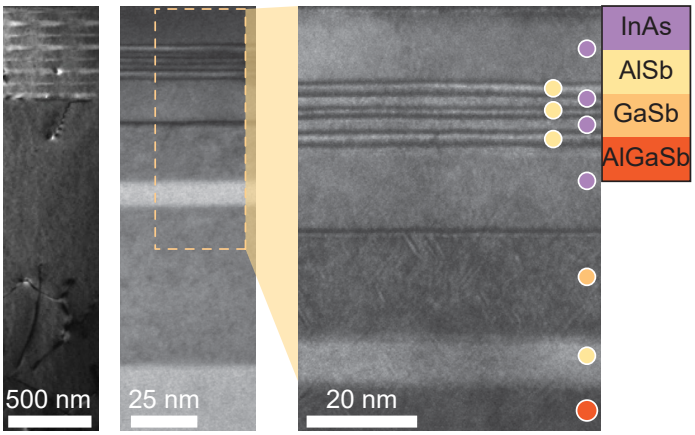

Figure 4.13: Cross-sectional TEM images at different scales of the XPH 2318 wafer. The coloured dots denote the corresponding materials in the legend. Images provided with permission from Dr Richard Beanland, University of Warwick.

Fabrication Flow

The latest fabrication process of the new design is illustrated in figure 4.14(a)-(l) and the corresponding processing photos of a 20-μm gate size device in a preliminary fabrication are shown in figure 4.15(a)-(i) for each step.

To start with, after pre-cleaning in figure 4.14(a), as represented in figure 4.14(b), in order to create a compact contact, a unique bilayer photoresist with S1813/PMMA is first spin-coated on top of the sample. Due to the dissimilar sensitivity to UV light, during the lithography process of S1813 removal step, only the top S1813 layer is exposed and removed, while the PMMA beneath is left untouched. The retained PMMA is left in place until the lift-off of the source-drain pattern step, where it is developed with the S1813 resist, ensuring that the deposited metal on top of the gate stack breaks off to form the compact contact. Then the first LW exposure defines the gate stack pattern and alignment fiducials. In practice, for a better alignment, metal markers are easier to identify than etched patterns in terms of image contrast etc., so an initial layer with alignment fiducials is first patterned and deposited with Au/Ti by thermal evaporation before the whole processing.

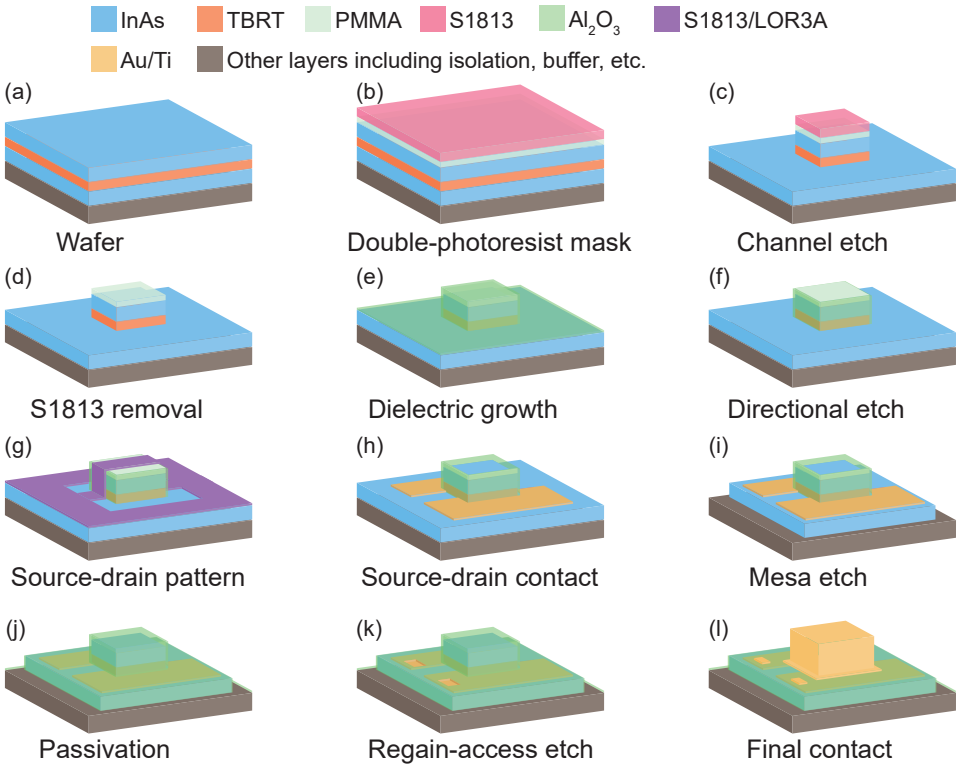

Figure 4.14: Schematic representation of the improved design with fabrication steps. (a) Wafer preparation. (b) Mask-photoresist spin-coating. (c) Channel etching. (d) Top S1813 photoresist removal. (e) Dielectric growth. (f) Anisotropic etching of the dielectric. (g) Source-drain contact patterning. (h) Source-drain contact formation. (i) Mesa etch. (j) Passivation layer deposition. (k) Regain access etch. (l) Final contact formation.

Following the self-aligned consecutive ICP etching of PMMA and the gate stack, the InAs layer in the channel is reached as shown in figure 4.14(c). The PMMA etching can also be traced and well controlled like the channel etching, as shown in figure 4.16(a)-(b). As photoresist coatings are capping the epi-layers, the fingerprints of transparent coatings are superimposed on the beginning of the usual reflectance curve for each wafer design, such that the etching depth or removal can be calculated by counting the ripples (fingerprints) before a well-known TBRT reflectance position. This applies to the next etching of the hardened S1813 layer as well. The PMMA

etching is calibrated with pre-test etch-through on a sample with the relevant layers to acquire information on reflectance fingerprints that are unique to each layer. The comparison between the reflectance from an etch-through and the simulation is plotted in figure 4.16. The slight difference can be explained by the thickness variation between the simulation and the actual thickness of the PMMA used. Given that the top S1813 has been etched multiple times by this point, and thus, been chemically changed, an ICP-monitored etching is performed to remove the hardened S1813 to ensure a successful photoresist stripping in the next step. A full exposure using optical mask lithography is carried out afterwards to remove the S1813 left on the chip for preparation of the next dielectric growth, leaving a PMMA capped gate stack as shown in figure 4.14(d).

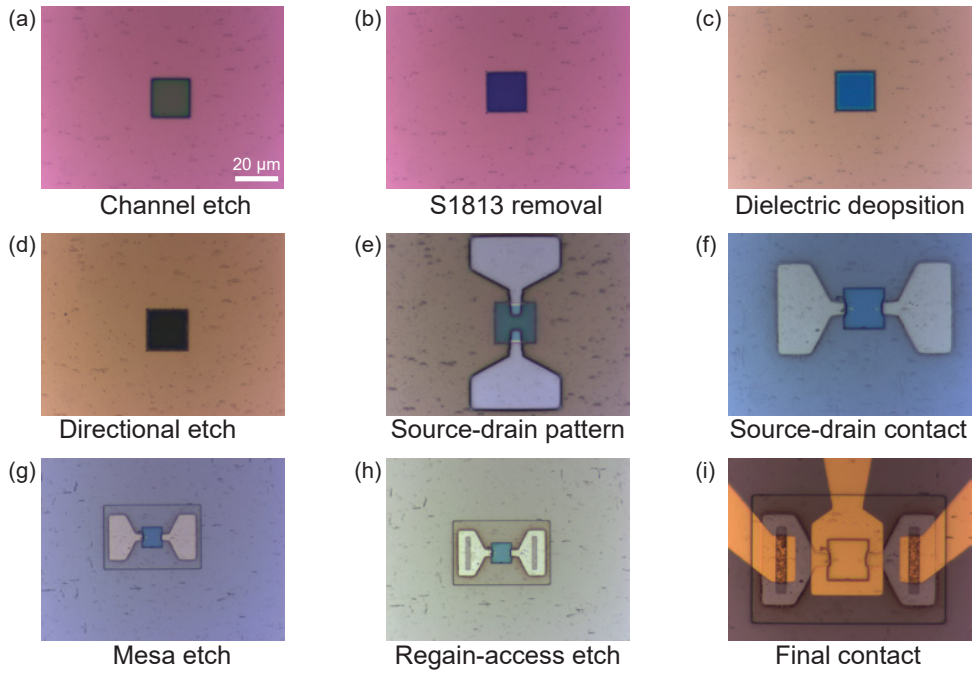

Figure 4.15: Fabrication photos of each step using the improved design. (a) Channel etching. (b) Top S1813 photoresist removal. (c) Dielectric growth. (d) Anisotropic etching of the dielectric. (e) Source-drain contact patterning. (f) Source-drain contact formation. (g) Mesa etch. (h) Regain access etch. (i) Final contact formation.

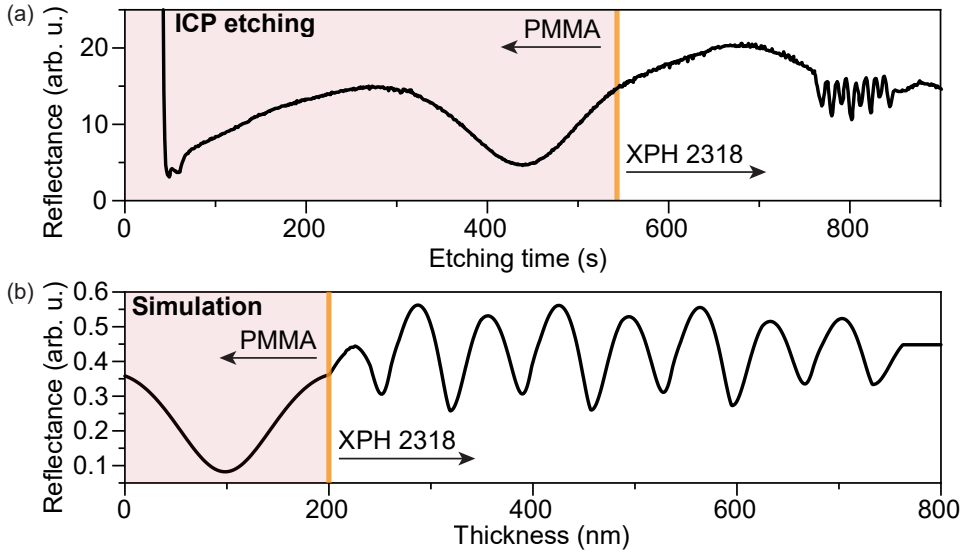

Figure 4.16: (a) Etching reflectance from the etch-through of PMMA/XPH 2318 structure. (b) The simulation for the structure in (a). Reflectance fingerprints of the PMMA are superimposed on the beginning of the known wafer reflectance. The orange line marks the stop line for end-point detection. Double layers of PMMA were used in the etching with a nominal thickness of 200 nm, which was used for the simulation in (b). The deviation between (a) and (b) can be attributed to the variation of PMMA thickness which depends on the spinning and baking conditions. The initial drop was caused by focus adjustment at the beginning of the etch due to signal issue.

Onwards fabrication continues with dielectric growth. For the thermal stability of the PMMA capping layer during the ALD process, the dielectric deposition is completed at 80 °C with 400 cycles of alumina. The deposited alumina layer forms an all-around gate stack isolation to protect exposed sidewalls as well as a separating layer between the gate stack and the subsequent source-drain contact, as shown in figure 4.14(e). The next self-aligned etching is accomplished by RIE, which is preferred for its high anisotropy etching. However, for a controllable etching, the ICP with end-point detection is used here with adjusted RF/ICP power ratio for a better directionality. The reference for the alumina etching is plotted in figure

4.17(a). For this case, the end-point detection spot was placed on a designated region that has the same condition as the gate stack which has the structure of Al_2O_3/PMMA/XPH 2318. The other alternative here is that, as simulated in figure 4.17(b), the spot can be focused on the channel region, which is supposed to give a better signal identification.

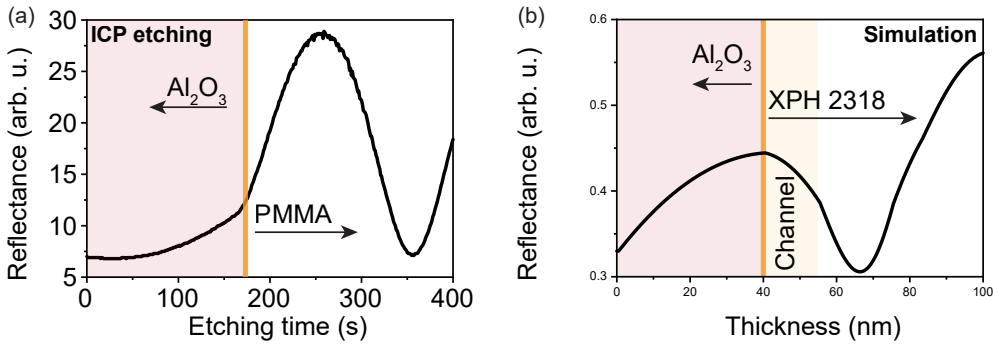

Figure 4.17: (a) Etching reflectance from the etching of Al_2O_3/PMMA/XPH 2318 structure. (b) The simulation for Al_2O_3/Channel/Other layers beneath the channel from XPH 2318 structure. The orange line marks the stop line for end-point detection. 40-nm thick Al_2O_3 was used for the simulation.

The entire alumina layer except for the covering on sidewalls will be removed by the directional etching, building an isolation layer separating the gate stack and the source-drain contact deposited later, as shown in figure 4.14(f). Next comes the patterning of the compact contact as shown in figure 4.14(g). The wrap-around contact is changed to a finger-shaped contact for an easier break-off process in the following lift-off process, as shown in the processing photo in figure 4.15(e). Following is the metallisation with a thin layer of 20 nm of Ti and 60 nm of gold, and the left PMMA capping will break off together with unexposed S1813/LOR 3A to form the close-enough source-drain contact to the gate stack, as shown in figure 4.15(h). Thanks to the capping PMMA layer, the alignment in this step is mismatch tolerant as any misalignment in any direction will be counteracted by the gate cap's break-off, while keeping the remained metal contact staying close to the gate stack.

The rest of the processing follows mesa etch, passivation, regain-access etch and final contacts, as shown in figure 4.14(i)-(l). The mesa construction is built on the third $Al_{0.7}Ga_{0.3}Sb$ layer of isolation units. The passivation layer used here is 30 nm of alumina grown at 150 °C, as usual. Regain-access etching is carried out by RIE. The final contact is deposited with 20 nm of Ti and 160 nm of gold. In addition, the overlapping gate design in figure 4.14(l) is a boon to the electrostatic control over the gate stack for memory operation.

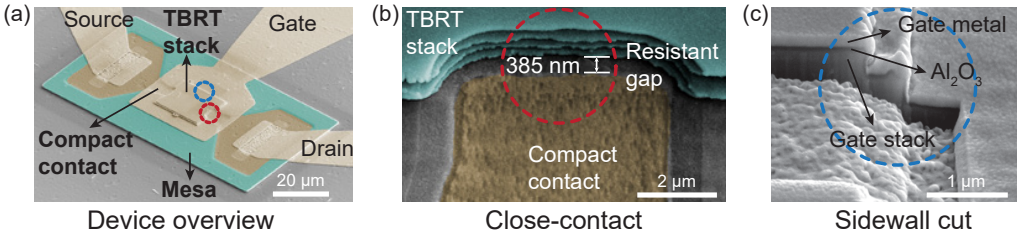

Figure 4.18: (a) Colourized SEM image of an as-fabricated device with improved design. The green region outlines the exposed mesa area while gold colour indicates the top electrodes. (b) Zoom-in image of the red-circled region in (a), showing a reduced resistive gap of around 380 nm. The green filled region denotes the TBRT stack while the gold colour region represents the compact source-drain contact. (c) The FIB cut at gate stack region as indicated by blue circle in (a), showing a high-quality sidewall.

Figure 4.18(a) shows an SEM image of a device fabricated by the improved design. With such design, as shown in the SEM measurement, the resistive gap between the gate stack and the source-drain contact is reduced to 385 nm, less than 500 nm, at least ten times smaller than the earlier design. In principle, with further optimisations, a gap dimension that is only determined by the dielectric thickness (down to a few nanometres) which is controllable by the number of ALD cycles, can be achieved. Following that, with appropriate annealing treatment for the source-drain contact, the total elimination of the resistive gap is feasible by controlled diffusion. The layered structure in figure 4.18(b) in the vicinity of close-

contact in SEM is likely to be caused by the etching from developer and remover used in the processing which can be avoided by further optimisation. The FIB cut in figure 4.18(c) suggests high quality sidewalls at the gate stack and the conformal coverage of alumina over the stack. Despite the step-like edge, which is likely attributable to an improper development process, the presence of a nearly vertical sidewall confirms the directional nature of the ICP etching. With further optimisation, precise control at the nanometre scale appears achievable. However, detailed quantitative characterisation of lateral etching is required to enable such fine control.

4.3 Channel Etching with End-Point Detection

The core concept of ULTRA**RAM**™ is the TBRT, which contains five atomically-thin layers with a total thickness of 10.2 nm. The whole fabrication process, regardless of design or version, lies in a top-down method, necessitating a fine control of the etching of the TBRT, a nano-scale carving. Efforts to develop a reliable and precise etching process including investigations on wafer design, etching recipe and end-point detection improvement have been made.

4.3.1 Layout Design Improvement

A thicker channel is a straightforward way to allow longer time for end-point control. Additional layers to shift the reflectance, or different layer arrangements to modify the reflectance signal shape, provide the other ways to a more reliable and precise end-point processing.

However, due to the critical thickness limitation in MBE growth, material options left for layer configuration adjustment are restricted to AlSb, GaSb, etc. After multiple attempts, the combination of a thickened 15 nm channel with an extra AlSb/GaSb layer between the channel and isolation layers was found to deliver a better reference signal in terms of etching monitoring. This new layer configuration

was implemented the growth of XPH 2093 wafer. Simulations in figure 4.19 shows the reflectance comparison between XPH 1896 wafer and XPH 2093 wafer. In comparison to XPH 1896 where the channel is grown on top of isolation layers directly. The insertion of AlSb/GaSb reduces the chance of over-etching by a flatter reflectance slope following the channel signal. Also, the reference to stop upon is shifted from a rise following a trough, with no clear change in the signal at the channel beginning, to a position where the channel starts right after an apparent peak that makes it easier to identify during the etching process.

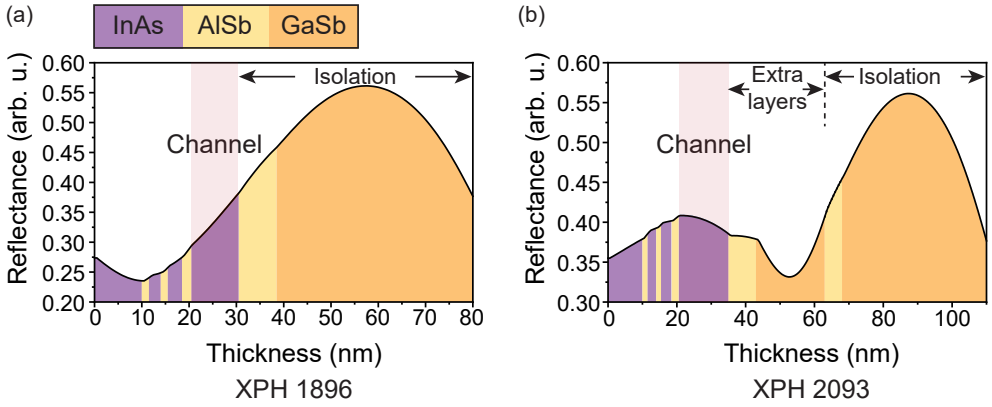

Figure 4.19: Simulations of reflectance for (a) XPH 1896 and (b) XPH 2093 at 670 nm wavelength, the additional AlSb/GaSb can be used as an over-etching signal.

4.3.2 Recipe Optimisation

On the other hand, endeavours to refine the recipe also involves the slowing down of the etching to allow a wider time window for etching control. The earlier version of the ICP recipe using CH_4 leaves a time window of $\sim$ 35 seconds and requires ashing between etchings to reveal the vanished signal due to the deposition of polymer caused by the use of CH_4, as shown in figure 4.20.

With optimisations on gas ratio and monitoring parameters, the CH_4-based recipe offers more details in the reflectance curve. In terms of the alignment between the simulation and ICP etchings, the best channel etching of ICP reflectances

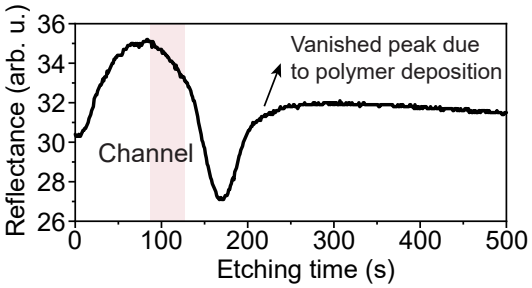

Figure 4.20: A reflectance from the mesa etching of a fabrication on XPH 1823 wafer using CH_4-based recipe, showing a narrow time window and the flattened signal caused by the polymer deposition.

in XPH 2213 with the optimized CH_4-based recipe shows a nice matching with the simulation, as plotted in figure 4.21(a)-(b). In particular, three visible kinks corresponding to three AlSb layers in the TBRT were observed as indicated by red arrows. The first dip within the ten seconds in the beginning is due to the plasma stabilisation. However, the quality of such detailed features varies across different etchings and wafers. Moreover, as indicated by the reflectance in figure 4.21(a), the time window for the channel region is limited to 30 seconds due to the fast etching rate.

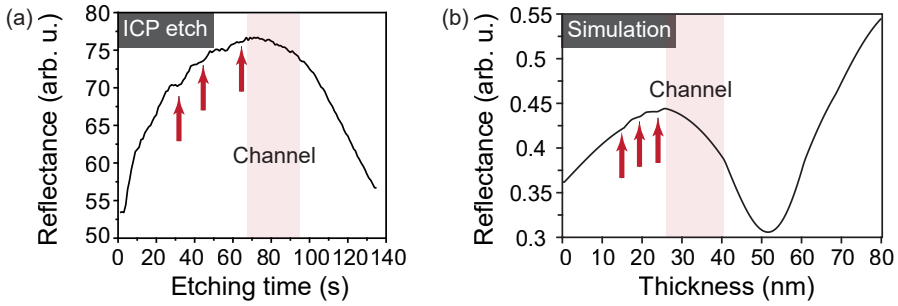

Figure 4.21: (a) Best ICP etching from XPH 2213 using the optimised recipe, showing clear matching with (b) unique features of the TBRT in the simulation.

To address the issue with CH_4-based recipes, a lower-power version, BCl_3-based recipe is developed. The monitoring parameters in the recipe are also well-adjusted

to ensure enough sampling points in signal detection and an appropriate smoothing rate for live plotting of the reflectance curve during the etching. The switching from CH_4 etchant to halide-based ones also helps to remove the blister-bubble issue that comes from the polymer deposition due to CH_4 usage. The reflectance of an etch-through on XPH 2318 wafer with well-defined and slowed-down version recipe is plotted as a function of etching time in figure 4.22(a), displaying a time window of 50 seconds for the channel etching control. The comparison to figure 4.22(b) reveals the agreement between the experiment data and the simulation result. The slow etching rate at the beginning of the curve can be attributed to a combination of factors, including material, etchant and surface oxidation. The recipe delivers a slower etching rate at the thin TBRT region and a faster removal at the thick mesa region as expected.

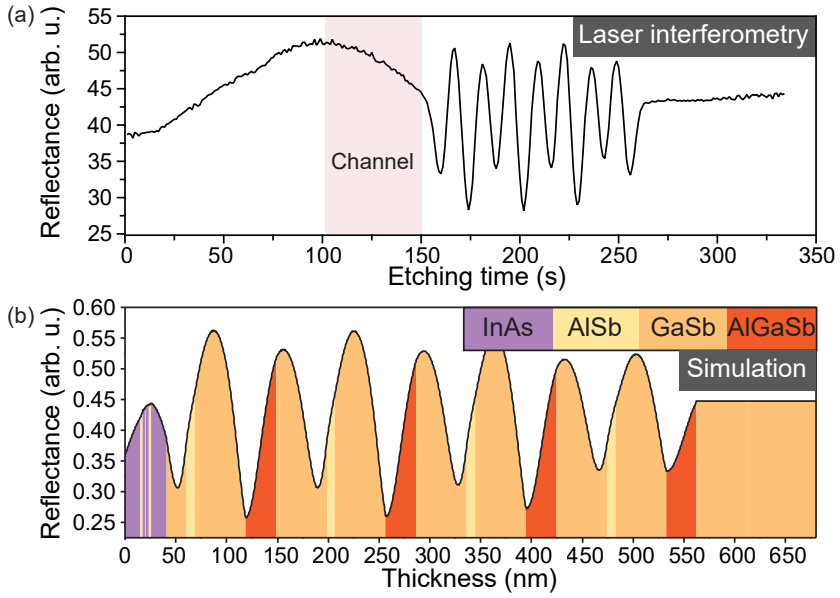

Figure 4.22: (a) Etch-through reflectance of XPH 2318 wafer with its reference to (b) the simulation of the same structure.

As the mesa etching is much easier to stop upon, this can be further accelerated by a higher ICP power in the recipe to save time. Although the bottom simulation shows peak-to-peak consistency with the etch-through reflectance, it is worth noting

that dissimilar etching rates for various material can stretch the reflectance curve, in comparison to the simulation where intensity is plotted as a function of thickness rather than time, and the etching rate is not taken into account. Therefore, the simulation is a general guidance and should be carefully used as a reference for end-point detection. An etch-through of the entire structure is always recommended to align the reflectance data to the simulation by identifying unique fingerprint features, as slight changes in growth thickness can shift the end-point signal away from its simulated reference point.

4.3.3 Laser Interferometry with Shorter Wavelengths

In terms of etching control, the best practice of a 15-nm channel etching with the optimised BCl_3-based recipe achieves a 6.5-nm thick InAs at the source-drain contact region after the whole processing, with 8.5 nm of InAs channel over-etched. Although the later regain-access etching also plays a role in etching the InAs, given that the non-halide recipes have better selectivity over InAs than BCl_3-based ones, the channel over-etching constitutes the primary reason for the reduced InAs thickness. Further improvement for the etching control is necessary. However, the improvement pertaining to end-point monitoring remains restricted, as the recipe optimisation, in principle, does not affect the fingerprint signals of a fixed growth structure. In addition, power and gas flow cannot be reduced indefinitely to slow down the etching, otherwise, the plasma becomes hard to strike. Moreover, alternatives in the material and modifications of the thickness regarding the epitaxial design provide limited changes to refractive index n and extinction coefficient k which shapes the final reflectance curve. Therefore, no substantial improvement can be achieved within these fields. However, on the interferometry side, in terms of the laser interference, the wavelength used can be considered to introduce further variability in the reflectance.

Modelling of the reflectance of XPH 2213 was performed with various wavelengths to find viable candidates. As plotted in figure 4.23(a)-(h), the channel is

highlighted by the colour filled area. In comparison to the original 670-nm modelling, shorter wavelengths offer a better resolution at the TBRT region with a higher intensity contrast between InAs well and AlSb barrier while longer wavelengths flatten fine features of the TBRT. The 206-nm simulation suggests the best shape for end-point detection with sharp transitions before and after the channel layer. However, due to the availability of products from suppliers, three wavelengths are finally used to carry out the experimental verification with the same etching recipe. Significant enhancement has been achieved, benefiting from shorter wavelengths. This is verified by multiple batches of wafer, including XPH 2093, XPH 2213 and XPH 2318, with three wavelengths of 340 nm, 365 nm and 405 nm.

As shown in figure 4.24(a)-(d), experimentally obtained etch-through reflectances are in agreement with the simulations of three wavelengths listed in figure 4.23(c)-(e). 2093_A and 2093_B are two pieces taken from same wafer XPH 2093. As seen in the highlighted channel region, all three wavelengths selected produce more detailed features that are useful for end-point detection at the TBRT region compared to the 670 nm one, with 365 nm delivering the optimal among them all with a kink signal right before the channel start, making it much easier to identify and to stop the etching accurately. Moreover, it has the best reproducibility across various runs.

The 405 nm also presents more features at the TBRT region in XPH 2093 sample, but the fingerprint feature before the channel falls short of repeatability in other samples. In addition, it shows no sign of over-etching for post-channel region as no visible contrast of signal intensity or features between InAs channel and the following layer. This applies to the 365 nm wavelength as well on XPH 2318 sample where the signal transition gets smeared from the channel to the underneath layer. This can be accounted for the non-uniformity across the wafer. The black lines in all plots are collected from a broadband detector, and are used as chamber plasma reference where the initial abrupt increase indicates the etching start, all signals before this point seen in the figure are caused by focus and positioning adjustments. The final sharp drop marks the ending of the plasma.

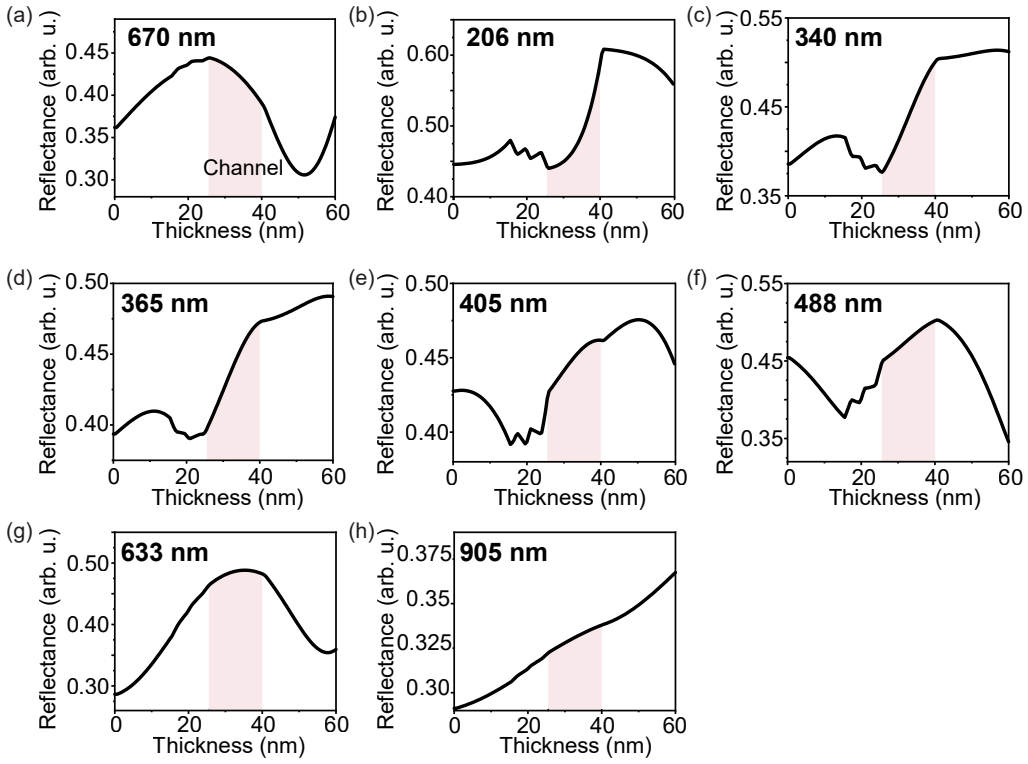

Figure 4.23: Simulated etching reflectances for various wavelengths at room temperature with focus on the TBRT region, fine features of InAs/AlSb get a higher contrast at shorter wavelengths while being flattened at longer wavelengths. The red boxes mark the channel region in each curve. (a) 670 nm. (b) 206 nm. (c) 340 nm. (d) 365 nm. (e) 405 nm. (f) 488 nm. (g) 633 nm. (h) 905 nm.

As-etched reflectances show slight deviations from the simulation result, this can be attributed to the n and k values used in simulation that are taken from bulk properties, and the actual temperature during the etching is not taken into consideration in the modelling. Although the 206 nm simulation shows the optimal curve for end-point detection, due to the limited availability of short-wavelength products, further fine-control may be achieved by atomic layer etching which works analogously to ALD in a self-limited way to ensure a monolayer removal per cycle.

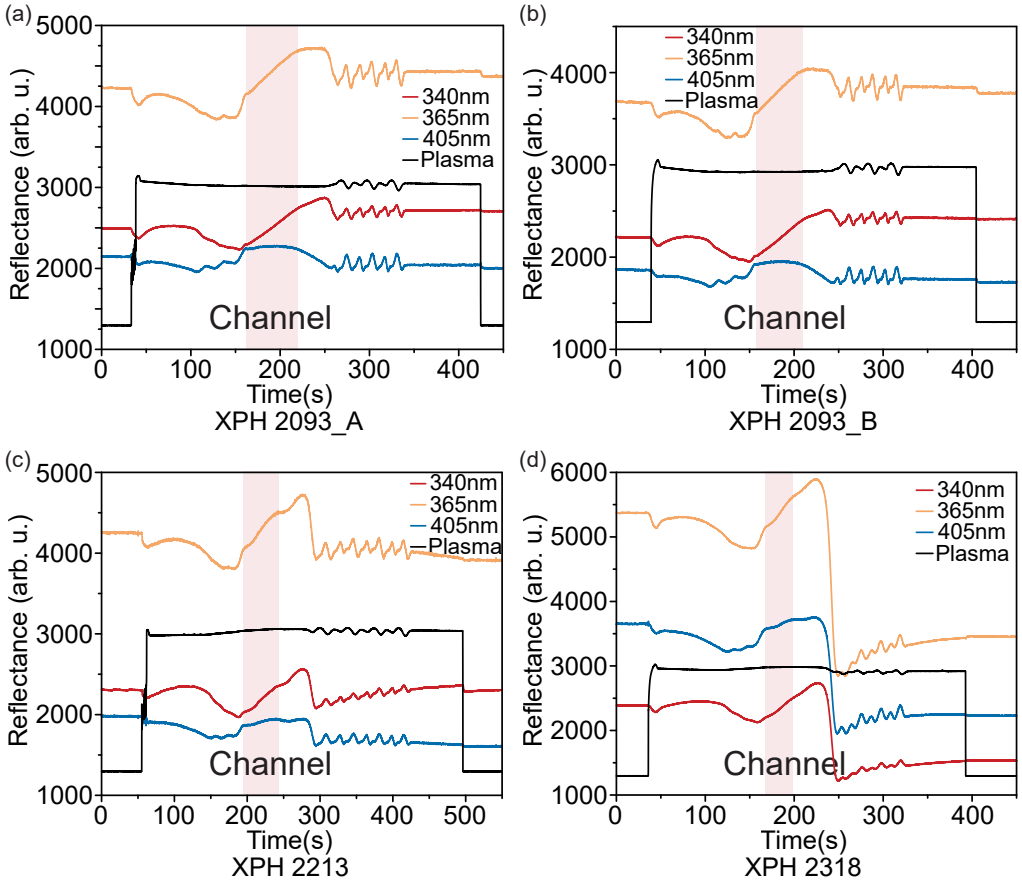

Figure 4.24: Reflectances of etch-through on samples including (a) XPH 2093_A, (b) XPH 2093_B, (c) XPH 2213 and (d) XPH 2318 acquired using three wavelengths of 340 nm, 365 nm and 405 nm, showing the consistency of fingerprint features of the TBRT across various designs. The etching and data collection were done in collaboration with David Cornwell, LayTec AG.

4.4 Summary

To conclude, three process designs ranging from the first version using wet-etching to the final improved self-aligned design with compact geometry, are discussed and demonstrated with experimental fabrications, showing the shortcomings solved and the advancements gained in each design. The primary challenges in the fabrication, the scalability and the resistive gap between the gate stack and the source-drain

contact in the previous design, have been addressed with the new device design. The proposed compact design has achieved the desired target in terms of scalability, indicating that the fabrication of smaller devices down to 50 nm gate dimension is, in principle, feasible. In particular, the size of the resistive gap has been significantly reduced by a factor of 10. With appropriate post-processing, such as controlled annealing, the contact gap could be eliminated, thereby meeting the objective of achieving a zero-gap contact. The progress in the optimisation of channel etching, which is the most critical part of ULTRA**RAM**™ processing, is also analysed. Laser interferometry at shorter wavelengths shows remarkable enhancement to the end-point detection of ICP etching by a more accurate identification of the TBRT during etching, offering further mitigation of the channel over-etching during the fabrication. In addition to precise control of vertical etching, accurate management of lateral etching is also necessary for scaled devices. For future nanometre-scale devices, further optimisation of the bias to ICP power ratio, along with a quantitative analysis of the vertical-to-lateral etch rate ratio, is essential to achieve this objective.

Chapter 5

Measurements and Characterisation

5.1 Gate Leakage

The earlier devices fabricated using a self-aligned design exhibited significant gate-to-source/drain leakage, primarily attributable to the use of Nb as a hard mask. This issue may be understood from two perspectives: the choice of material and the associated fabrication process.

In terms of the material, the refractory metal Nb was selected as the hard mask due to its high etch resistance. However, the Nb film is deposited via a sputtering process which is much more energetic than the usual thermal evaporation. Consequently, sputtered Nb atoms with high kinetic energy penetrate into the alumina layer during sputtering process, causing diffusion into dielectric layer. Therefore, the insulating property of alumina is compromised by the metallic Nb diffusion, resulting in the leakage across the dielectric layer. The diffusion phenomenon is proved by the XPS measurement that was carried out on a Nb/Al_2O_3/XPH 1896 structure where the Nb was deposited by sputtering, analogous to the gate structure in a memory device. To perform the XPS measurement, the sample surface is etched by a certain amount using Ar milling,

and then XPS signals from the fresh top surface are collected and analysed. After enough etches, the element distribution in the etching direction can be plotted and analysed. Figure 5.1 displays the result of the XPS measurement with corresponding III-V elements. Oscillating lines of Ga and Al reflect the periodic structure in the epitaxial design of XPH 1896, as shown in table 5.1. As the percentage is indicative, the Nb box (red) and the Al + O box (yellow) are placed to roughly outline the distribution of the three elements. The Nb region shows a clear overlapping with the combination of Al and O regions (an approximation of the Al_2O_3 layer). At the peak percentage of Al and O, there is $\sim 5\%$ Nb detected, indicating the diffusion of Nb into the dielectric layer. Furthermore, the high diffusion coefficient of Nb in Al_2O_3 worsens the phenomenon [243].

Layer	Material	Nominal thickness (nm)	Thickness measured by TEM (nm)
Floating gate	InAs	10	9.7
Quantum barrier	AlSb	1.8	1.9
Quantum well	InAs	2.4	2.4
Quantum barrier	AlSb	1.2	1.7
Quantum well	InAs	3.0	3.1
Quantum barrier	AlSb	1.8	2.7
Channel	InAs (n-type 5×10^{18})	10	10
Isolation × 3	AlSb	8	8.3
	GaSb	50	49.3
	$Al_{0.7}Ga_{0.3}Sb$	30	35
	GaSb	50	50
2-step buffer	GaSb (hot)	540	-
	GaSb (cold)	1400	-
Nucleation	AlSb	5.2	-
Substrate	Si	~ 380 μm	-

Table 5.1: Detailed layout of XPH 1896 wafer for ULTRA**RAM**™, illustrating the typical design of the floating gate, TBRT, channel, and isolation structures.

In addition to the problematic gate material, the other important factor to consider is the processing design in the fabrication. Being a hard mask, the Nb film is inevitably etched at the following self-aligned etching step, as a result, some Nb is sputtered around and susceptible to be re-deposited onto sidewalls of the exposed gate stack, creating a conductive path for leakage current. The mesa etching step

exacerbates the etching-induced sputtering issue due to the higher power and the longer etching time for the aforementioned exposed gate regions (as depicted in figure 4.10(d)).

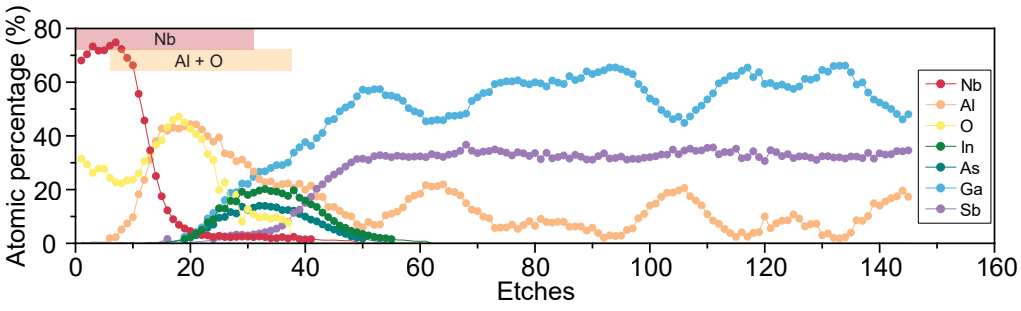

Figure 5.1: XPS characterisation of multiple layers on XPH 1896 wafer. XPS signals are collected and analysed after each run of milling down with a certain amount of the thickness. The Nb signal overlapping with the combination of Al and O signals shows the diffusion of Nb into dielectric layer. The percentage numbers are indicative.

In order to solve the issue, Ta metal was first tested due to the higher etch resistance than Nb. Despite the sputtering process, the Ta/Al_2O_3/Au structure displays much less leakage. Due to its smaller diffusion coefficient than that of Nb in Al_2O_3 (Cr: $10^{-15.59}$ cm^2/s < Nb: $10^{-13.50}$ cm^2/s at 1200 °C) [243], thermal evaporated Cr was then characterised on same structure in an effort to protect the vulnerable alumina layer from sputtering process. As shown in figure 5.2 (a)-(b), both tests show a leakage of 10^{-10} A magnitude, three orders of magnitude smaller than the mA leakage seen in the Nb device. However, Ta and Cr cannot erase the etching-induced sputtering issue as both are conductive metals, and create a leakage path once are sputtered. A Si_3N_4 layer is proposed later as an extra protective layer for the metal hard mask to mitigate the sputtering issue, but this consequently produces a more intricate fabrication process. Additionally, Cr is not compatible with the current process in terms of etching recipe. Therefore, in the end, as a final resolution, Si_3N_4 is adopted as the etching mask material.

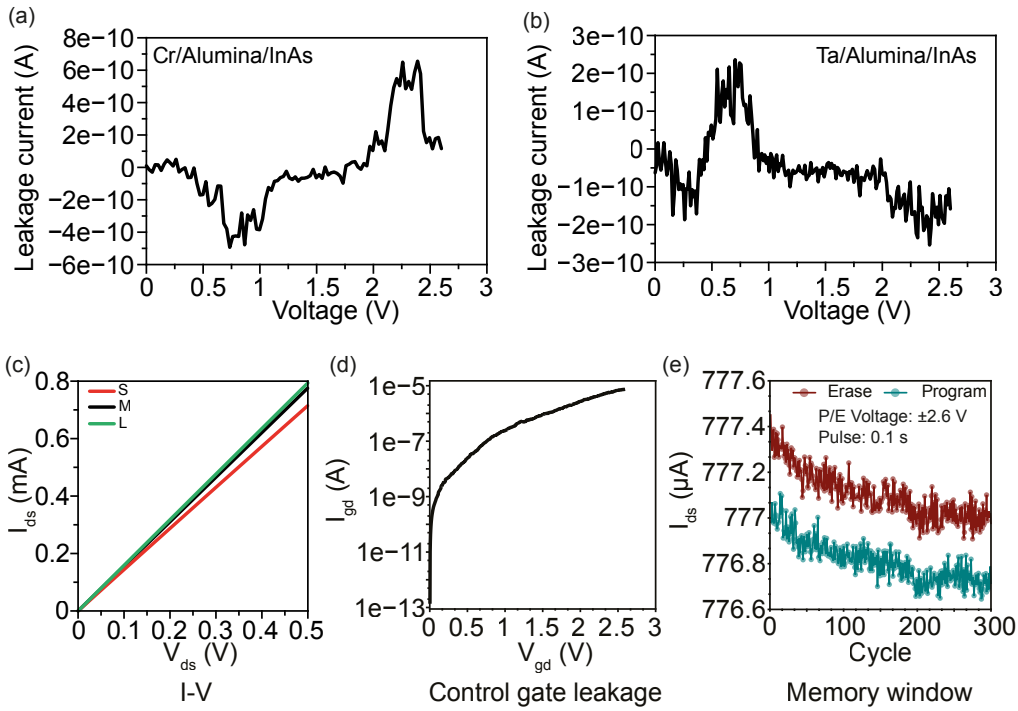

Figure 5.2: Leakage test for (a) Cr and (b) Ta on alumina. (c) I-V curves for devices with gate size S, M and L. The gate dimensions of S, M and L are 10.5 × 20 μm^2, 25 × 30.5 μm^2 and 30 × 56 μm^2, respectively. (d) Gate leakage and (e) memory window plot of the memory device of size M in (a) from XPH 1823 using Si_3N_4 as mask for gate definition.

Figure 5.2 (c)-(e) shows the basic measurements for the devices fabricated using the Si_3N_4 mask on XPH 1823 wafer. The linear I-V curves from devices of different gate dimensions in figure 5.2 (c) were measured between the source and drain contacts of the device, which show a good conductivity of the channel layer underneath the TBRT region. Figure 5.2 (d) shows a gate leakage of ~ 7 μA between the control gate and the drain, representing the solving of the leakage issue. The endurance test in figure 5.2 (e) was performed through repeated cycles of erasing and programming the memory cell, as described in section 3.3.1. The difference of channel current between the erased state and the programmed state establishes an endurance memory window, which is a fundamental characterisation in a memory

test. A memory window of 0.3 μA was observed in the as-fabricated memory cell. Detailed discussion on the memory characterisation can be found in the following section 5.3.

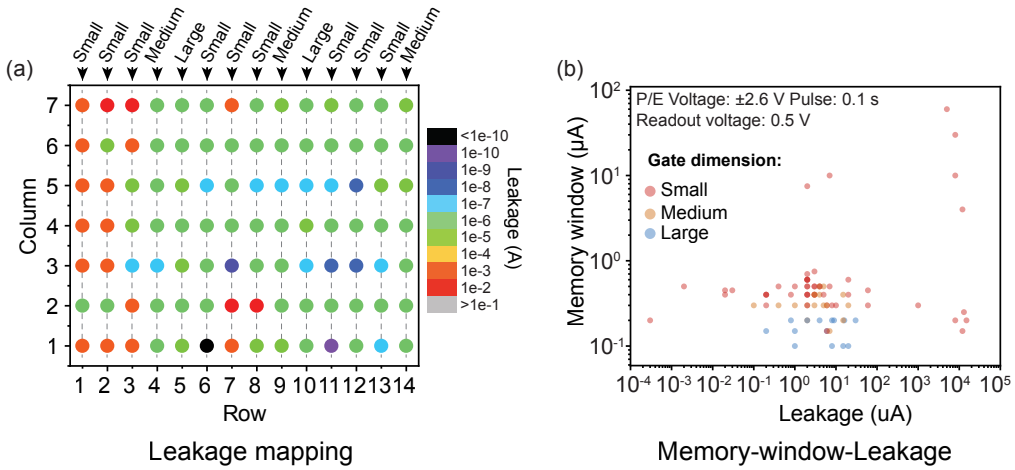

Figure 5.3: (a) Leakage mapping and (b) memory-window-leakage plot of the fabrication from XPH 1823 using Si_3N_4 as mask for gate definition. The device positioned in the first column of the sixth row is non-functional.

To further investigate the fabrication process, figure 5.3 presents a comprehensive characterisation of all devices from the fabrication batch on the XPH 1823 wafer, including a mapping analysis of gate leakage and the corresponding memory window sizes as a function of leakage, for a total of 98 devices on the chip. The leakage mapping illustrates the spatial distribution of leakage current across the physical layout of the as-fabricated chip. Notably, over 80% of the devices exhibit gate leakage currents below 10 μA, representing a substantial improvement compared to the mA leakage previously observed in devices employing Nb hard masks in earlier process flows [7]. This marked enhancement in performance is a critical step towards meeting the stringent yield requirement of 99.9% necessary for large-scale industrial production. All devices exhibiting significant leakage possess small gate dimensions and are predominantly located within the top three rows of the chip. This location-dependent leakage is plausibly attributed to misalignment during the

optical mask lithography step. The gate dimensions of small, medium and large devices are 10.5 × 20 μm^2, 25 × 30.5 μm^2 and 30 × 56 μm^2, respectively. The array of devices exhibits a periodic variation in dimensions-small, small, small, medium, and large-as indicated in figure 5.3(a). The memory-window-leakage graph in 5.3(b) shows a statistical analysis, conveying that regardless of location, the size of the memory window is irrelevant to the gate leakage. The colours red, yellow, and blue correspond to small, medium, and large gate sizes, respectively. Most of non-leaky devices have an endurance memory window around 0.5 μA. This is due to the poor channel readout performance, and the limited gate control over floating gate, as illustrated in figure 4.9(h) in section 4.2.2, due to the device architecture, the gate contact area is intentionally designed to be smaller than the gate stack, a consequence of alignment requirements during the patterning process. Thus, a reduced memory window was obtained. These two concerns led to the latest improved self-aligned design. In summary, the characterisation of those memory devices implies the successful approach of Si_3N_4 as etching mask for addressing the gate leakage issue.

5.2 Channel Characterisation

The channel is integral to the device construction, exerting a significant impact on the memory performance. Before diving into the memory characterisation, fundamental characterisations of the channel including circular transfer length method (CTLM) and basic tests on field-effect transistors built on the memory channel are performed to gain insights into channel properties.

5.2.1 Transfer Length Method

The earlier growth designs, ranging from XPH 1823 to XPH 2213, utilise a doped channel with an increased carrier concentration of 5 × 10^{18} cm^{-3}, in an attempt to enhance the channel conductivity. However, the effect seems to be balanced

out by the deterioration of mobility introduced by the doping-induced scattering, as a slightly better conductivity is observed in the undoped InAs. Figure 5.4(a) shows a basic I-V measurement for two samples with same geometry and etching conditions, with one from the doped XPH 2213 wafer and the other from the undoped XPH 2318 wafer. The current measured in doped devices has exhibited variation across different samples and wafers; however, no significant differences have been identified. Consequently, the suboptimal performance is unlikely to stem from fabrication issues, suggesting that the undoped devices demonstrate better performance. Therefore, in the latest design, an undoped InAs layer is used for the channel layer. On the other hand, doping is essential for providing carriers. Modulation-doped InAs/AlSb structures have demonstrated that barrier doping can enhance the carrier mobility of the two-dimensional electron gas within the quantum well [244, 245]. This doping strategy may be further optimised for TBRT by introducing dopants in the layer either above or beneath the channel.

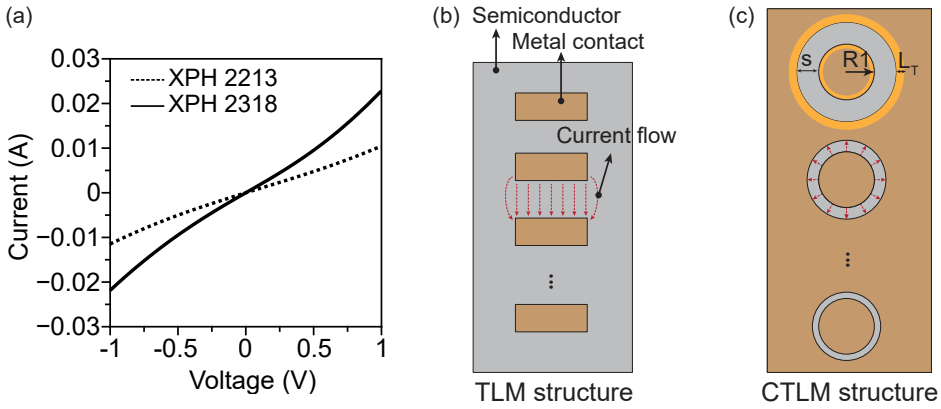

Figure 5.4: (a) I-V measurement of samples from XPH 2213 and XPH 2318 under same etching conditions. (b) TLM and (c) CTLM structure. The dashed arrows in (a) shows the current crowding which causes the current to flow through other sides of the pad. s, the gap between inner electrode and common ground; R1, the radius of inner electrode; L_T, the transfer length.

It's worthwhile to check the channel quality, as well as the contact condition

of the metal/undoped InAs interface in the first place. The transfer length method (TLM) is commonly implemented to characterise the contact resistance of a metal/semiconductor interface. In this regard, it is of significance to investigate the contact resistance of memory devices. However, the TLM has a drawback that current flow is not limited only to the region defined by the gap, as it also spreads to other sides of the contact pad due to current crowding as illustrated in figure 5.4(b). For that reason, the CTLM is introduced to account for the geometrical issue. The CTLM structure is illustrated in figure 5.4(c) where multiple circles of same dimension (R1) are separated from an electric field by various ring-shaped gap spacings (s).

For the measurement of CTLM, standard I-V sweeps are carried out between each circular pad and the common pad. Each sweep contributes to a point on the total resistance - spacing gap curve. In order to be analysed similarly for the linear fitting, CTLM data points need to be divided by correction factors (c) to avoid underestimation [246] using

$$c = \frac{R1}{s} ln \frac{(R1 + s)}{R1} \,, \tag{5.1}$$

while the total resistance obtained with a multiple linear regression analysis can be expressed as

$$R_{tot} = \frac{R_S}{2\pi R1} s + 2R_C \,, \tag{5.2}$$

where R_{tot} is the total resistance across two measured pads, R_S is the sheet resistance of InAs between two metal pads and R_C is the contact resistance of the metal/InAs interface. Thus,

$$R_S = slope \times 2\pi R1 \,, \tag{5.3}$$

where the slope is given by

$$slope = \frac{\Delta R}{\Delta Spacing} \,. \tag{5.4}$$

The intercept of the ordinate is $2R_C$ while the intercept of the abscissa is $2L_T$. L_T, the transfer length. It refers to the distance over which injected charge carriers travel

through the semiconductor material beneath the contact before being collected into the contact.

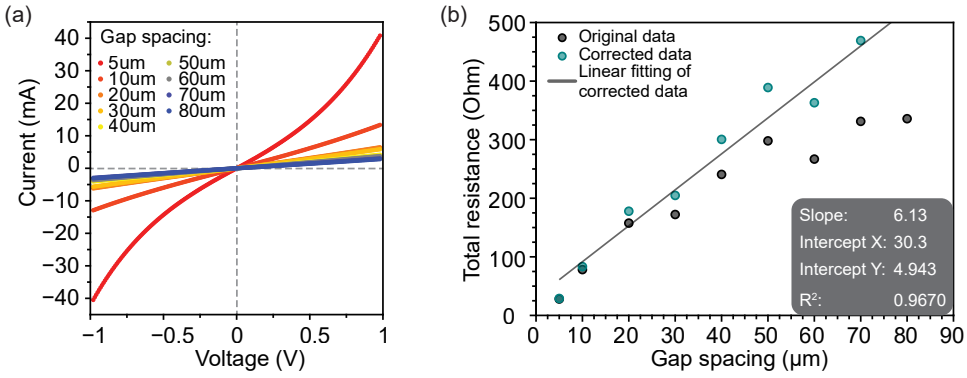

Figure 5.5: (a) I-V measurement data for CTLM and (b) CTLM analysis of XPH 2318 wafer. Nine different gap spacings were used in the characterisation. R^2, the coefficient of determination.

To prepare the CTLM measurement, the entire sample was first etched down to the InAs channel layer using ICP etching with reflectance-based end-point monitoring, followed by a single lithography process with the CTLM pattern. Contact pads were formed afterwards by a thermal evaporation of an Au/Ti film onto the exposed InAs layer. On each data point, two probes were placed onto a circular pad and the common pad at the same time to perform an I-V sweep. The slope of each sweep was then extracted from the linear fitting of the output curve to contribute to a data point in the CTLM characterisation. Figure 5.5(a)-(b) displays the I-V sweeps for the CTLM measurement and the analysis of the TLM test on XPH 2318 wafer. A total of nine different gaps (5, 10, 20, 30, 40, 50, 60, 70, 80 μm) were used with corresponding inner pads (R1 = 75 μm). Most I-V sweeps are generally linear, indicating an Ohmic contact at the metal/InAs interface. Based on the calculation above, derived values including R_S, R_C, specific contact resistance

$$\rho_C = R_S \times {L_T}^2 \tag{5.5}$$

and L_T from figure 5.4(c) of the as-etched undoped InAs channel of XPH 2318 are

listed in table 5.2. The coefficient of determination for the linear fitting is 0.9670.

R_S ($\Omega\square$)	R_C (Ω)	ρ_C (Ωcm^2)	L_T (μm)
2888.69	15.15	1.76×10^{-4}	2.47

Table 5.2: Measured values extracted from the CTLM measurement of XPH 2318 wafer.

5.2.2 Field-Effect Transistor Measurement

To further examine the properties of the ULTRA**RAM**™ channel layer, a MOSFET device featuring the InAs channel was devised, and multiple electrical tests were performed to observe the channel modulation which is of vital importance for the readout operation of ULTRA**RAM**™ memory. A MOSFET is a three-terminal semiconductor device that modulates current flow via an applied electric field. It comprises source, drain, and gate terminals, with a conducting channel—InAs in this case—formed between the source and drain within a semiconductor substrate. The gate is electrically insulated from the channel by a thin oxide layer. The objective of this study is to characterise the linear region of the as-fabricated MOSFET. As illustrated in figure 2.3 (section 2.1.1), MOSFET characterisation involves measuring the transfer characteristics, in which the channel current is recorded as a function of the gate voltage at a constant channel bias, as well as the output characteristics, where channel current measured as a function of channel bias at a fixed gate voltage. These measurements are fundamental for analysing device performance and extracting key operational parameters.

MOSFET devices were built on the InAs channel using the same wafer for ULTRA**RAM**™ fabrication. Fabrication steps are illustrated in figure 5.6 (a). The wafer was first etched down to the channel, then contact pads were patterned and metallised. Mesa etching was used to isolate different devices. Next, a 30-nm alumina layer was grown by ALD to conformally cover the exposed surface, acting

as a dielectric layer and a passivation layer. Finally, contacts and control gates were placed at the last step.

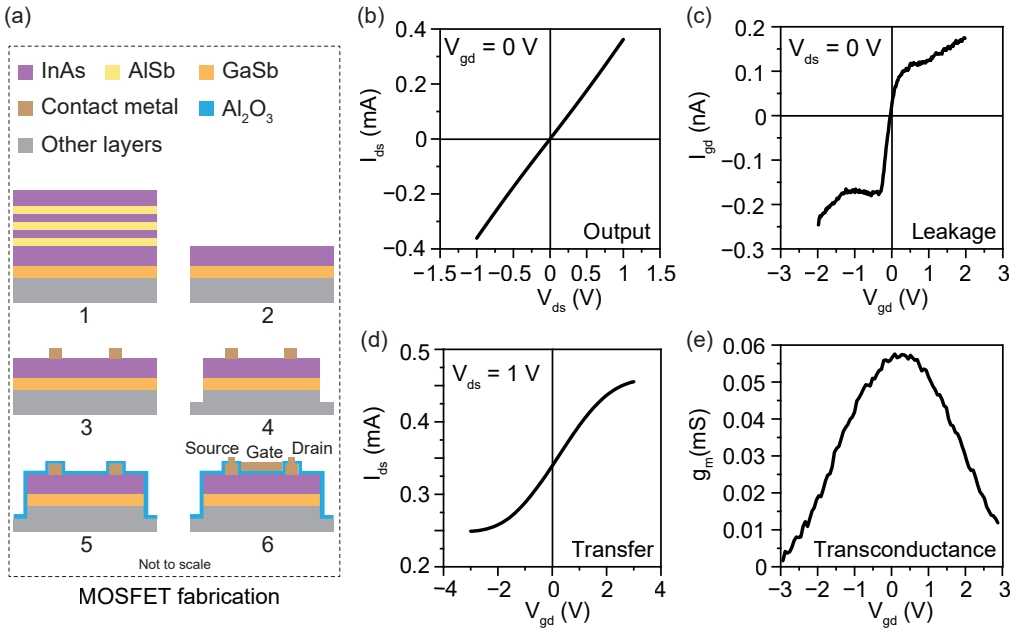

Figure 5.6: (a) MOSFET fabrication steps for the measurement. (b) Output, (c) leakage, (d) transfer and (e) transconductance characteristics of a MOSFET device fabricated on XPH 2318. I_{ds}, the source-drain current; V_{ds}, the source-drain voltage; I_{gd}, the gate-drain current; V_{ds}, the gate-drain voltage; g_m, the transconductance.

MOSFET characterisation of XPH 2318, as well as basic I-V and leakage tests are shown in figure 5.6(b)-(e). The linear I-V curve of mA scale at 1 V channel bias seen from the drain characteristic in figure 5.6(b) suggests good contact condition at the metal/channel interface, and an acceptable channel conductivity. However, given that the operating voltage for ULTRA**RAM**™ readout is less than 1 V at zero gate bias, the gate voltage and channel bias were set to zero and ± 1 V only. As a result, the full MOSFET characteristics, including operation in the saturation regime, were not observed. The sub-nA leakage at 2 V gate voltage as shown in the leakage plot in figure 5.6(c) exhibits high insulation from the ALD dielectric layer. The I_{ds}-V_{gd} sweeping at 1 V channel bias in figure 5.6(d) shows a clear modulation

of the channel conductivity by the control gate within the memory operation voltage range, from -2.6 V to +2.6 V, indicating a successful gate electrostatic control over the InAs channel. This is crucial for the memory readout operation as the binary discrimination can only be sensed when the channel is modulated by the charges in the floating gate. The higher current observed with an increasing positive gate bias in the transfer characteristic indicates the n-type nature of the InAs channel. Non-zero current at zero bias confirms the normally-on property of the InAs channel of ULTRA**RAM**™. Transconductance is a critical parameter that describes a MOSFET's ability to convert a change in gate voltage into a change in drain current. It is defined as the ratio of I_{ds} to V_{gd}, under the condition of a constant drain-source voltage. This current-to-voltage ratio is commonly referred to as the gain. The formula for deriving the transconductance from a MOSFET measurement is given by

$$g_m = \frac{\Delta I_{ds}}{\Delta V_{gd}}\,. \tag{5.6}$$

Figure 5.6(e) includes the transconductance plot of the MOSFET extracted from figure 5.6(d), showing a maximum of 0.0576 mS at zero gate bias, the normalised transconductance corresponds to 2.22 mS/mm. For comparison, a typical InAs FET exhibits a transconductance in the range of 400 to 1000 mS/mm with values up to 1600 mS/mm reported for 5 nm-thick InAs HEMTs [247]. Though modulation is achieved in the MOSFET structure, the transconductance remains small, which implies a limited gate control over the channel region. Another contributing factor is that the etched InAs in the channel does not achieve the same electrical quality as the unprocessed material.

5.3 Memory Characterisation

Following the confirmation of the channel modulation, measurements on memory devices can be carried out. In this section, the measurement set-up and tests of memory devices fabricated from three batches of wafers are analysed and discussed,

establishing the correlation between the device performance and the processing design.

5.3.1 Measurement Configuration

All memory tests were carried out on a probe station connected to an SMU. Figure 5.7 presents the configuration and connection of a typical set-up used for memory characterisation, a three-terminal measurement configuration. During the memory test, the programming and erasing pulses are sent to the gate terminal via one SMU channel while the readout signal is sensed via the drain terminal using the other channel of the SMU. The source terminal is used as the common ground for both SMU channels during the measurement. All measurements carried out on the probe station were performed at room temperature under strong illumination from the microscope lamp of the probe station.

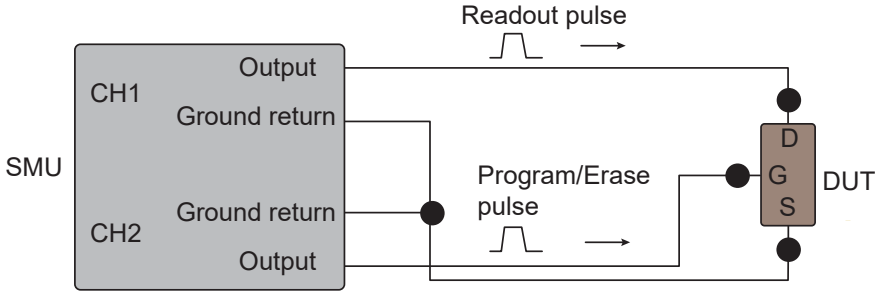

Figure 5.7: Schematic of measurement connections in a typical ULTRA**RAM**™ memory test. CHn, the measurement channel; G, the device gate terminal; D, the device drain terminal; S, the device source terminal; DUT, the device under test.

Endurance and retention are the workhorses of memory performance characterisation. Both are implemented through a sequence of pulses using an SMU. For a typical endurance test, repeating program-read-erase-read operations are carried out to test the endurance of a memory device by measuring the size of the memory window over many cycles. Retention test works similarly to endurance but only includes a single programming or erasing pulse at the beginning, followed

by continuous readout pulses over time, to examine the difference in channel current between two binary states.

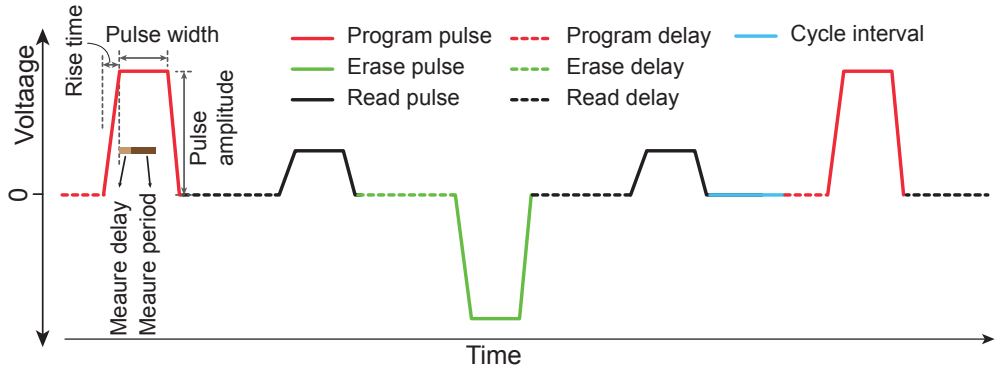

Figure 5.8: Pulse pattern for memory endurance characterisation, showing a multiple-level pulse waveform.

The sequential bias pattern used in ULTRA**RAM**™ memory for the endurance test is illustrated in figure 5.8: a multi-level pulse waveform that consists of multiple voltage segments and simultaneous voltage-current measurements. Key test parameters include pulse amplitude, pulse timing parameters (delay, rise time, pulse width, fall time) and measure parameters (delay and measure period). A cycle interval segment is inserted between cycles to address thermal concerns in extremely long endurance test. For ULTRA**RAM**™, a faster transition time and shorter pulse width are preferred. The pulse amplitude refers to the pulse height required to program and erase the memory cell. For the circumstance of ULTRA**RAM**™ memory, this value is usually set to 2.6 V. The pulse width is the time a voltage is held on for programming, erasing or readout operation. A longer pulse width facilitates greater energy delivery to the device and allows for improved signal settling, thereby enabling more stable and reliable measurements. In contrast, a shorter pulse width is advantageous for capturing rapid transitions and transient behaviours; however, it poses challenges in terms of measurement accuracy, particularly due to limited settling time. The measurement delay ensures the current/voltage measuring takes place after the pulse becomes stable and before

the pulse vanishes. The measurement duration is defined by the number of cycles and power frequency. A longer duration provides a higher accuracy, but the entire duration is limited to the length of the pulse width. Customisable pulse parameters allow the flexibility in pulse pattern to meet the device-specific requirements for ULTRA**RAM**™ memory tests.

In the experimental evaluation of memory behaviour of ULTRA**RAM**™, a series of voltage pulses were applied to the device using an SMU. The memory test protocol involved applying square wave pulses with various pulses. For the program/erase operations, the pulse width was varied between 0.005 s and 0.1 s. Depending on the specific pulse period used, the duty cycle can be calculated using the relation

$$Duty\,cycle = \frac{Pulse\,width}{Pulse\,period}\,. \tag{5.7}$$

The pulse amplitudes were set to + 2.6 V for programming and - 2.6 V for erasing. For the readout operation, a longer pulse width was employed to ensure more accurate measurement of the channel current, with the read pulse amplitude fixed at 0.5 V. These parameters were chosen based on preliminary studies indicating optimal switching behaviour in this range, while minimizing device degradation. The response of the memory cell was monitored using the other channel of the SMU, recording changes in channel current as an indicator of memory state transition, enabling the construction of evaluation of memory endurance and retention characteristics.

5.3.2 Device from Batch XPH 1823

This batch of ULTRA**RAM**™ devices were fabricated on XPH 1823 wafer. As outlined in table 4.2 (section 4.2.2), the XPH 1823 wafer features the introduction of isolation layers. The self-aligned design with Si_3N_4 as the etching mask was used for the fabrication.

Figure 5.9(a) shows the static drain characteristic of a typical ULTRA**RAM**™ memory cell from XPH 1823 at room temperature, the linear I-V curve suggests the

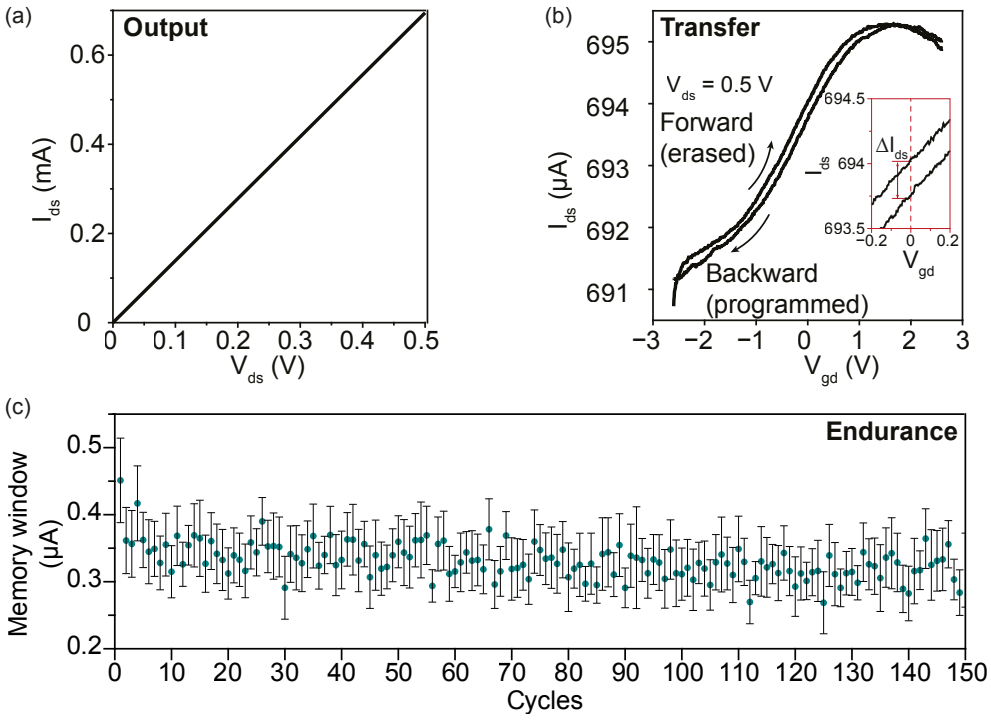

Figure 5.9: (a) Output, (b) transfer sweep and (c) endurance memory window as a function of switching cycles for devices from XPH 1823 with a readout voltage of 0.5 V. Endurance data are shown as mean ± standard error from multiple devices. The inset in (b) depicts the current discrepancy ΔI_{ds} at zero gate bias which implies the existence of charges stored in the floating gate for programmed status. I_{ds}, the source-drain current; V_{ds}, the source-drain voltage; V_{gd}, the gate-drain voltage.

channel conductivity and no contact barrier at the contact/channel interface, also a signal of a successful channel etching in the fabrication. The transfer curve at 0.5 V channel bias in figure 5.9(b) displays a clear hysteresis that represents the charge storage in floating gate. The inset in figure 5.9(b), the zoom-in of the hysteresis at zero gate bias, depicts a discrepancy of 0.25 μA at zero gate bias, indicating the change of channel conductivity induced by the charges stored in the floating gate. The number of electrons in the floating gate can be estimated by

$$N = \frac{C\Delta V}{q} \approx 9.9 \times 10^6 \,, \tag{5.8}$$

where C is the capacitance of the floating gate and ΔV is the threshold voltage change between two states, q is the elementary charge. A short endurance test of 150 cycles for multiple devices is plotted in figure 5.9(c) to show the memory window measured at 2.6 V programming/erasing voltage (negative bias for erasing and positive bias for programming). The memory window is defined by the difference between erased current and programmed current. Based on the calculation, the endurance window in the measured memory cell is around 0.35 μA in the endurance test. This is in agreement with the current discrepancy in the transfer characteristic in figure 5.9(b). The pulse time for memory operation measurement is 0.1 s and the readout bias is set to 0.5 V. Regarding reproducibility, the majority of devices on the chip exhibited memory characteristics.

To further explore the memory performance of the device, a group of varying voltages and pulse lengths were taken into consideration on another ULTRA**RAM**™ device. Up to 3 V operation voltage and down to 5 ms pulse width were applied to characterise the endurance response of the memory device. As described in figure 5.10, each red dot represents a single readout of the channel current from an erased status while each green dot represents a single readout current of the channel after a programming operation. The first three sessions were carried out with a decreasing pulse length for memory operation. As seen in figure 5.10, the size of the memory window shrinks over a shorter pulse width. The logic state discrimination can be sensed with a pulse time down to 5 ms at 2.6 V operation voltage. As the pulse width increases, the memory window generally broadens, up to a point of saturation. A longer pulse duration facilitates greater charge injection into, or removal from, the floating gate, thereby enhancing the shift in threshold voltage. This results in a more pronounced distinction between the programmed and erased states. Then an increased pulse voltage was applied to test the pulse amplitude response of the memory cell. In both pulse widths used, the higher programming voltage shows an improvement to the size of memory window but not significantly, for the instance of identical 0.1 s pulse width, the memory window is enlarged from 0.5 μA to 0.6 μA.

Despite multiple continuous tests, the memory cell remains functional after 2000 cycles, exhibiting a good endurance performance.

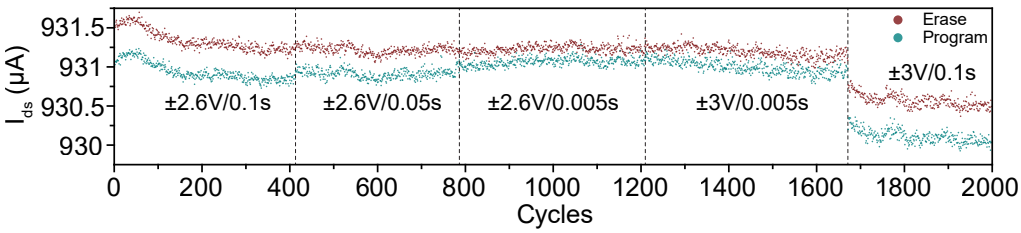

Figure 5.10: Endurance characterisation at various voltages and pulse widths of a device from fabrication on XPH 1823. I_{ds}, the source-drain current. ±2.6 V/0.1 s denotes the memory operation voltage ± 2.6 V with a pulse duration of 0.1 s, this convention applies similarly to all subsequent annotations. Readout voltage is set to 0.5 V for all measurements.

The small memory window is a common phenomenon across all devices on the chip, which can be attributed to leakage currents, i.e. alternative current paths between source and drain other than via the InAs channel, exacerbated by the parasitic resistance due to a resistive gap between the source/drain contact and the gate stack caused by the channel over-etching, thus a poor channel readout performance. The cross-section of the self-aligned device is illustrated in figure 5.11 (a), the inevitable over-etching causes the exposed channel to be thinner than that of gate stack, which creates a resistive gap between gate stack and contact metal.

The thinned-down InAs channel has a reduced conductivity and may experience further declining when thinned down to a few atomic layers where the quantum confinement introduces discrete energy levels such that limited electrons can contribute to the transport. This is supported by the observed larger memory window at a higher temperature where more thermal activated electrons give a higher channel conductivity, as shown in figure 5.11 (b). The over-etched channel encourages the channel current to flow through the underlying GaSb layer (hole current) during the memory read operation, which will respond oppositely to

electrons in the InAs channel with regard to the charges in the floating gate, competing with the InAs channel current (electron current), leading to a poor memory performance. In this self-aligned design, the scale of the resistive gap is comparable to the gate feature size, i.e. a few microns, significantly degrading the channel conductivity for memory readout operation.

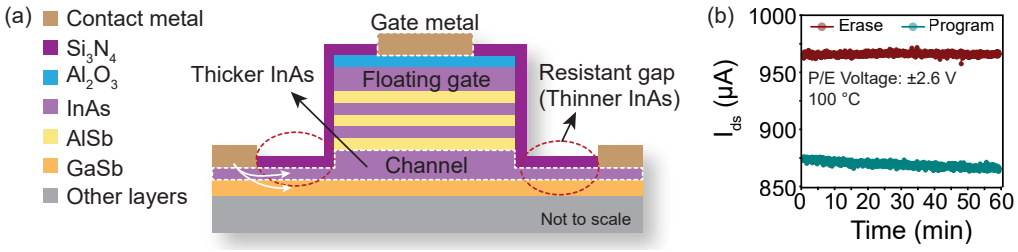

Figure 5.11: (a) Cross-section sketch of the device fabricated with self-aligned design, showing drawbacks of native resistive gap. The gate metal is smaller than the floating gate, so there is limited gate control over channel region. (b) Retention test at 100 °C. The P/E voltage was achieved by a half voltage scheme with + 1.3 V applied on the source and - 1.3 V applied on the drain. The high temperature measurement was performed by Charlie Senior and Max Walker Long.

A separate consideration is the concern of attaining adequate electrostatic integrity. The Si_3N_4 replacement of the problematic metal film as etching mask is not without drawbacks, as can be seen in the top of gate stack in figure 5.11, due to a re-gain access etching window, the late-placed control gate metal only covers a fraction of the floating gate region for the reason of alignment, which in return, imposes a limitation of the tight control of the channel charge by the gate during programming/erasing operation. If the shortcomings of over-etching and parasitic resistance are addressed, much better performance can be expected.

5.3.3 Device from Batch XPH 2213

This fabrication of ULTRA**RAM**™ with XPH 2213 wafer employed the same self-aligned process as in XPH 1823 where Si_3N_4 is used as the etching mask and the dielectric layer is 15-nm thick alumina. However, as concerns have been expressed about poor channel performance and an important goal is to maximise the memory window, an improved accurate etching with optimised CH_4-based recipe was implemented in this fabrication and the wafer features a thicker channel of 15 nm InAs to further enhance the over-etching tolerance. The thickness of the floating gate is increased to 15 nm as well to improve the charge storage capability for a greater memory window.

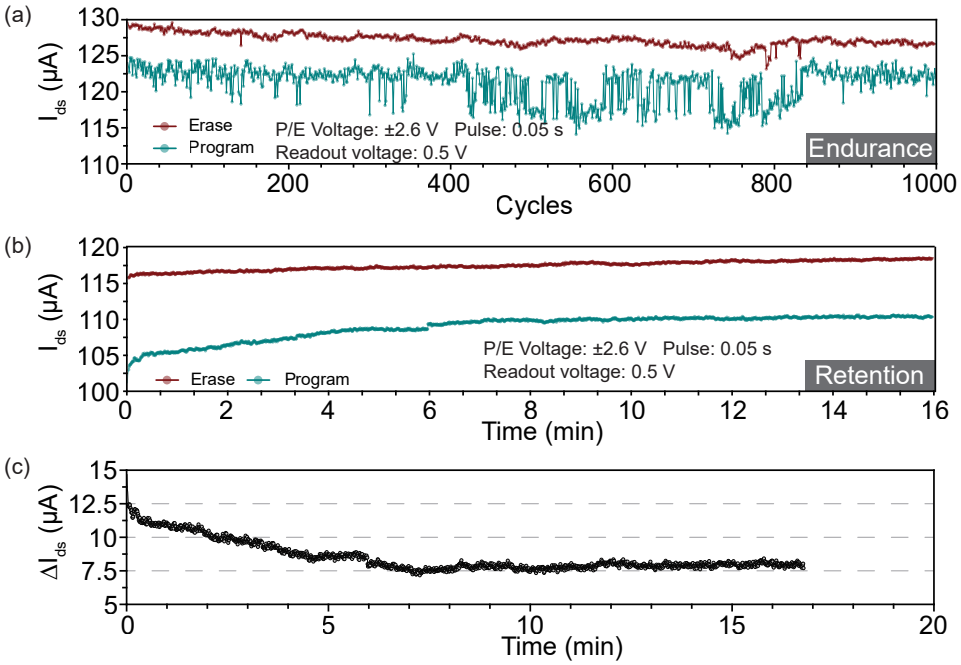

Figure 5.12: (a) Endurance, (b) retention and (c) retention memory window of a memory device from fabrication on XPH 2213, showing a stable memory window of 5 μA. I_{ds}, the source-drain current; ΔI_{ds}, the retention memory window.

Memory performance including endurance test and retention test are plotted in figure 5.12. Memory behaviour can be observed and a robust memory window

of 5 μA (10 times better than that of XPH 1823 fabrication) is achieved in both endurance and retention measurements. However, as shown in figure 5.12(a), the endurance measurement shows an unstable programmed current over the 1000 measured cycles. The observed 'bistable' programmed state current is likely attributable to partial programming in some cycles—where only a portion of the floating gate area is charged—resulting in higher current levels, while other cycles exhibit full-area programming, corresponding to lower currents. An alternative explanation may involve the occupation of discrete energy levels within the floating gate across different cycles. It is unlikely that this behaviour arises from defects in the TBRT structure, as these would be expected to trap varying amounts of charge in a non-systematic manner during the memory programming process. The retention test in figure 5.12(b) exhibits an initial window of around 13 μA, but the programmed current gradually increases until a stable memory window around 7.5 μA is eventually reached after six minutes and remains stable within the rest of the measurement period, and no sign of degradation seen from the projection as seen in figure 5.12(c). The initial rising of the programmed current is likely to be charges coming out of the traps in the TBRT. The programmed status is stable for retention because it is only programmed once, not multiple times, i.e. no variation in the programming area because there is just one program operation. As a comparison, the erased status is relatively stable in both endurance and retention characterisation. The memory test pulse was set to 2.6 V and 0.1 s for programming and erasing operations while 0.5 V for readout measurements.

Underpinning the phenomenal improvement of memory window compared to the previous fabrication on XPH 1823 are two major optimisations, the enhanced channel performance where the over-etching issue is dramatically mitigated by the refined recipe and the increased thickness, and the thickened floating gate that contributes to the charge storage capability for memory operation. The thicker channel benefits from reduced surface scattering and enhanced carrier mobility, thereby exhibiting better conductivity than the over-etched, thinner counterparts.

To conclude, the combination of the well-defined channel and the thickened floating gate confer the development with the potential to deliver a boosted memory performance.

5.3.4 Device from Batch XPH 2318

Fabrication of ULTRA**RAM**™ on XPH 2318 wafer features an undoped InAs channel to further improve the channel readout performance and Ta was brought back as gate metal to address the issue that restricts the electrostatic control over gate stack originating from the Si_3N_4 etching mask design. To avoid the etching-induced sputtering issue, double layers of S1813 were used and left on top of Ta film for protection (photoresist is easier to fit in the processing without incurring substantial changes than using Si_3N_4). A BCl_3-based recipe was used for both channel and mesa etching.

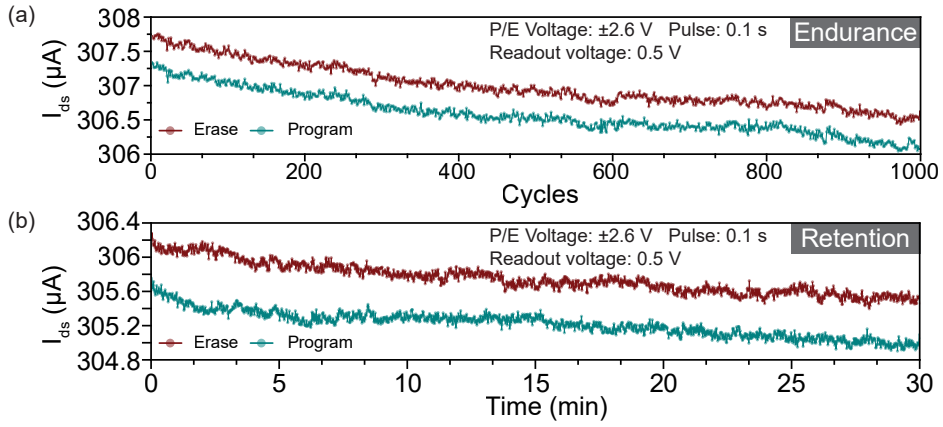

Figure 5.13: (a) Endurance and (b) retention characterisations of a memory device from fabrication on XPH 2318, showing a stable memory window of 0.5 μA. I_{ds}, the source-drain current.

As shown in figure 5.13, memory window is plotted as a function of number of cycles and time for endurance and retention, respectively. Both obtain a memory window of 0.5 μA at 2.6 V and 0.1 s pulse condition. However, the drift observed in the current for both programmed and erased states exceeds the memory window,

which may be attributed to poor contact which causes the degradation of metal-semiconductor interface over cycles or from the instrumentation drift. This could be improved by further investigation on the InAs/contact TEM characterisation and more tests for optimal contact metal. Instrumentation drift can be ruled out by testing known-good devices. The resistive gap accounts for the significantly smaller memory window compared with XPH 2213. In the optimal example of the etchings, a 6.5 nm thick InAs layer was obtained in the final ULTRA**RAM**™ memory device, leaving a removal of 8.5-nm InAs in comparison to its primitive thickness of 15 nm, making up a 57% over-etching. The over-etching of the channel outweighs all other factors in terms of the thickness loss. Despite the small window, no sign of deterioration of the performance within 1000 measured cycles for endurance and 30 minutes duration for retention test. The memory window is at same scale as in XPH 1823 meaning no significant improvement from the introduction of Ta gate metal, though the channel conductivity of the undoped channel is better than that of XPH 2213. Despite the general declining current magnitude over cycling or time, the absolute memory size remains same over the measurement period, showing relatively stable TBRT performance of ULTRA**RAM**™.

To briefly summarise, the devices fabricated on XPH 1823 demonstrated the successful implementation of all-dry-etching; however, they exhibited limited performance. In contrast, the memories from XPH 2213 showed an improved memory window, attributed to advancements in growth design and etching processes. The subsequent batch, XPH 2318, which features a Ta metal gate and an undoped channel, did not exhibit further performance enhancement. The obtained channel thickness suggests that this limitation is likely due to channel over-etching. Despite the limited device performance, the vertical mesa profile achieved in the as-fabricated devices demonstrates the feasibility of the all-dry etching method for scalable design. While not yet at the nanometre-scale target, this fabrication offers critical guidance for the development the final self-aligned architecture.

5.4 X-Ray Nano-Probe Technique Analysis

Outside of device improvement, it is worthwhile to turn attention to the analysis of the failure mechanism of heavily cycled devices. In this regard, XANES spectra are useful in the analysis of the oxidation state or coordination environment of elements in the sample. XRF is helpful in chemical and elemental analysis. XANES was carried out on three types of devices: fresh, partially cycled and cycled to failure, to find any defects related to the cycling of devices.

5.4.1 X-Ray Absorption Near Edge Structure

Figure 5.14 shows the set-up for X-ray nano-probe measurement. Apparatus such as monochromator need to be first tuned for the target energy of the beam, and all the parts need to be well calibrated prior to that energy change of the system. As shown in the schematic, a high brilliance beam coming out of an undulator first enters the double crystal monochromator, and then, after being collimated, passes through the chopper to facilitate measurement with a lock-in amplifier. In the next, the beam passes the collimator, and afterwards hit the sample which is wire-bonded to a through-hole mount. Finally, the beam goes into the ion chamber. The detectors were placed near the sample surface at a fixed angle to collect as much signal as possible, and filters were put in front of the detector to reduce background noise and prevent detector saturation for enhanced signal-to-noise ratio. The ion chamber used here is for the normalization of the transmitted beam signal. The lock-in amplifier is used to measure signals at the frequency of the chopper in the connection loop.

Figure 5.15 shows the XANES measurement results of three devices. The ordinate is intensity signal of In Kα sum value normalized by GaAs sum value, i.e. a representation of absorption. The abscissa is the scanning energy of beam. As seen from the figure, there is a similar intensity shape versus energy for all devices, with no impact or dependency found between cycling number and device performance.

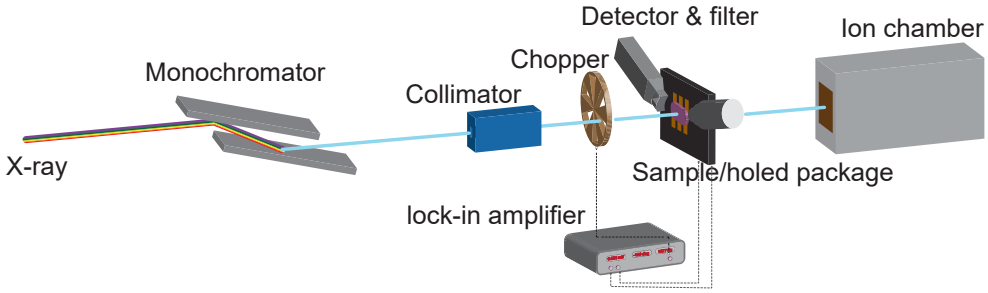

Figure 5.14: Illustration of set-up for XANES measurement.

In other words, there is no impact on In-related structure from cycling numbers.

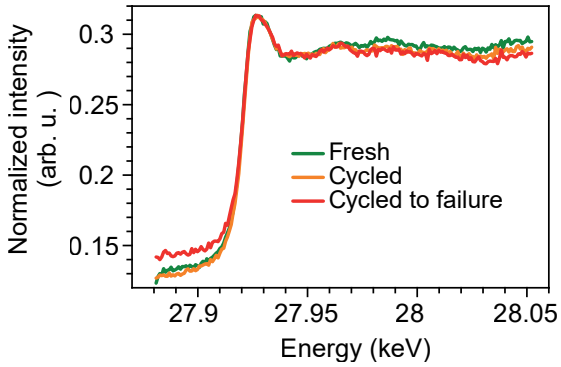

Figure 5.15: XANES curves of fresh, cycled and cycled to failure devices under In K-edge energy set-up, no evidence of In-related defects to cycling failure.

5.4.2 X-Ray Fluorescence

By setting the energy range around In Kα2 (24.00 keV), a mapping of region of interest for In could be obtained for nano-analysis, as seen in figure 5.16. The device (fresh) photo and selected area XRF intensity profile as a function of scanning position are shown in figure 5.16(a). The In distribution is same as the device design with the maximum at the gate area. However, a few paler spots were observed which suggests the lower In concentration in those spots. Figure 5.16(b) shows the photo and selected area XRF mapping for a cycled device. Similar to figure 5.16(a), a

few areas with lower concentration spots were found. Abscissa and ordinate are lateral and vertical scanning displacement, respectively, as indicated in the optical photos by red dash boxes. In general, XRF with In energy set-up shows certain areas with less In concentration across the device, which could be related to some inhomogeneity or defects of InAs in the TBRT region. To conclude, a systematic X-ray nano-probe measurement carried out on various memory devices implies that the cycling number has no significant effect on In-related defects in ULTRA**RAM**™ i.e. no strong correlation was found between the cycling-failure and In-related defects.

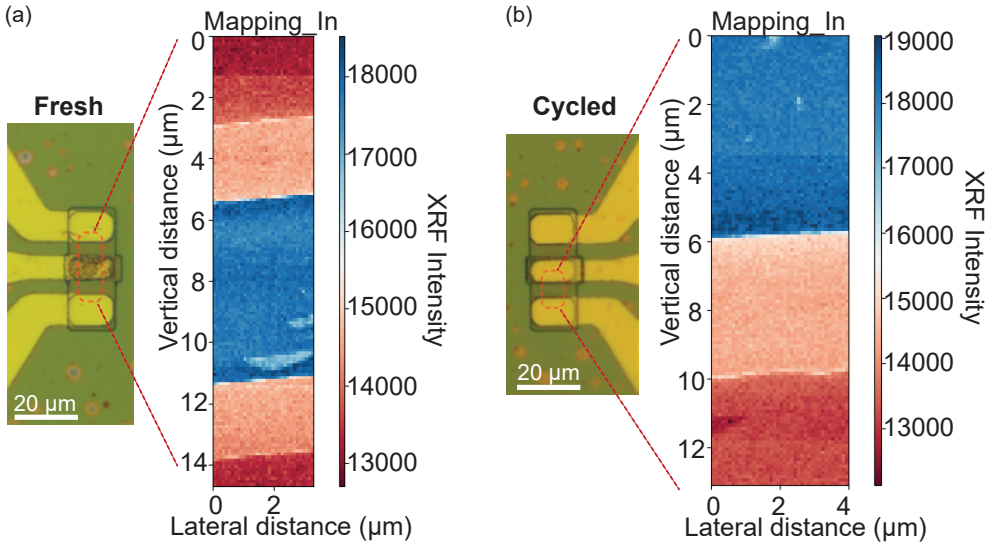

Figure 5.16: XRF mappings of (a) fresh and (b) cycled devices from XPH 1896 under In K-edge energy set-up, no significant difference is observed between fresh and cycled devices. The lateral and vertical distances correspond to the dimensions scanned, as indicated by dash boxes on device photos to the left of each mapping.

5.5 Summary

In summary, to start with, the gate leakage issue was addressed with Si_3N_4 as the gate definition hard mask. Then, fundamental characteristics were carried out to investigate the basics of the InAs channel layer for further memory analysis.

Next, three different batches of ULTRA**RAM**™ fabricated on various wafers, and with different processing conditions, are discussed and analysed. The enhanced endurance memory window of 10 μA in XPH 2213 provides a good testament to the significance of channel layer engineering. The Ta gate metal solution shows no significant improvement in terms of memory window though with improved channel conductivity that is likely from the undoped InAs channel. Since all measured devices were fabricated with a design that has a native resistive gap and the over-etching issue remains, the improved self-aligned design might be the next solution for the next step. Though not deployed yet, it is expected to deliver substantially better performance with further memory readout enhancement by the elimination of the resistive gap, improved gate electrostatic control and high scalability. In comparison to previous work employing wet etching, the successful demonstration of processing using all-dry etching represents a significant step towards a scalable design. Subsequent optimisation has led to further notable improvements. The proposed self-aligned device architecture, representing the first fully scalable design, exhibits promising results for the development of a scalable ULTRA**RAM**™ platform. Finally, an attempt to investigate the mechanism underlying cycling-induced failure, a systematic analysis of ULTRA**RAM**™ using X-ray techniques showed that no evidence of a link between cycling number and In-related defects, paving the way for the investigation of ULTRA**RAM**™ failure mechanisms.

Chapter 6

Simulation by nextnano

6.1 Background

Notwithstanding the negligible thickness discrepancy from target values for the ULTRA**RAM**™ structure, TEM imaging with atomic resolution shows a fluctuating interface where alloying or intermixing exists, i.e. an imperfect composition of the TBRT layers, as shown in figure 6.1. In figure 6.1(a), each dumbbell represents two atoms, as illustrated by the red and blue ball for In and As atom in the floating gate region. In figure 6.1(b), the compositional analysis shows that the incorporation of In and As into AlSb barrier layers in XPH 2093 wafer brings the barrier composition to quaternary rather than the intended binary AlSb. The In and As atom fraction are not zero at the peak of the Al and Sb signal, indicating that the interface is not chemically sharp as it's supposed to be, and the In and As distribution can be found across the entire barrier region in all three barriers measured. By way of comparison, the quantum well layers show a composition much closer to its original design, and it's almost pure InAs at the upper end of the layer in the growth direction. The intermixing is due to the exponential decay of the atoms that are diffusing from one layer to the next. It is the atoms that are on (or close to the surface) that are finding their way into the subsequent layers. A similar barrier alloying phenomenon was also found in the TBRT by scanning tunnelling microscopy in similar structures

[248] where memory layers were grown on a GaAs substrate.

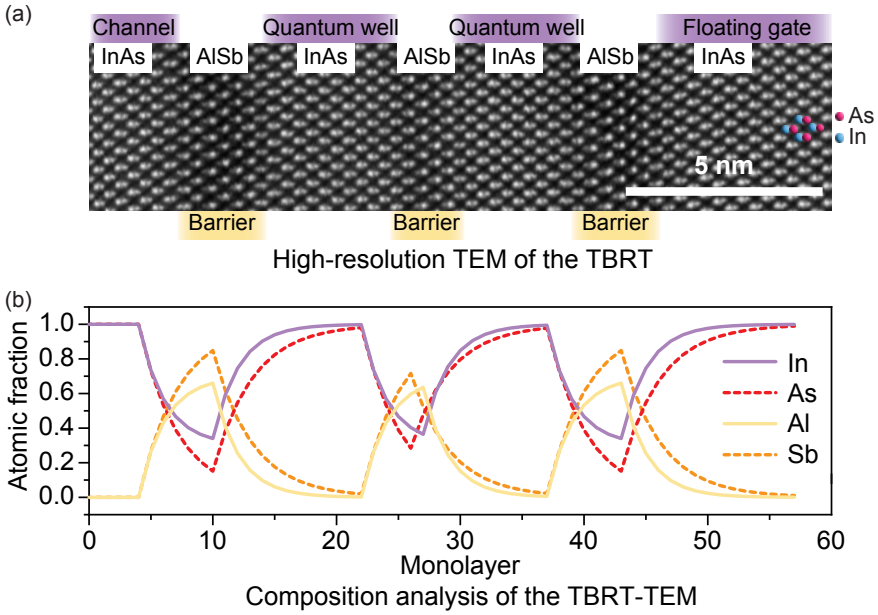

Figure 6.1: (a) Cross-section TEM of TBRT layers of XPH 2093 wafer. InAs and AlSb layers are denoted on top and bottom of the figure. (b) Compositional analysis of corresponding layers of the TBRT in (a). Images provided with permission from Professor Richard Beanland, University of Warwick.

A previous simulation investigating the effects of variations to the heterostructure layer thickness for TBRT performance optimization [214] , only considered changes to the thickness of the layers. However, the actual scenario of interface alloying, which is evidently more severe has not been previously considered. Therefore, in order to gain a comprehensive understanding of transport properties of the TBRT, a series of simulations were conducted to study the impact of TBRT alloying on the memory operation of ULTRA**RAM**™. The simulation was performed with one dimension along the growth direction for simplification. The nextano simulations are calculated from Schrödinger-Poisson solutions. The input parameters, including material properties sourced from databases and the defined device geometry, are first processed. This is followed by the calculation of band edges, and subsequently, a self-consistent solution of the Schrödinger and Poisson equations is carried out.

The MSB simulation is conducted under the assumption of a single-band model.

Defects from MBE growths such as threading dislocations are observed in ULTRA**RAM**™ wafers as well, as shown in the large scale TEMs in chapter 4, some of which reach to the top memory layers. However, since there will be hundreds of threading dislocations in every device, they have been proven to be inactive [249] in the tunnelling process and with limited effect to the transport performance [250]. A diode with InAs quantum well and AlSb barriers yielded a room-temperature peak to valley current ratio of 3.2, in spite of a 7.2% lattice mismatch between the InAs epilayers and the GaAs substrates, where a measured threading dislocation density of roughly 10^9 cm^{-2} exists [249]. Therefore, the material defects were not considered in these simulations.

In terms of the channel InAs layer, defects like interface roughness, layer thickness variation, threading dislocations, stacking fault, trap centres, inhomogeneities will detrimentally affect resonant tunnelling and the device performance of the heterostructure. It has been reported that the defects primarily influence the peak to valley ratio for resonant tunnelling diode device [251]. R. Magno et al. reported these defects act as a parallel current path that doubles the valley current compared to a defect free device [252]. In particular, a high concentration of As on Al antisite defects which are formed during the growth of the AlAs-like interface can increase the electron concentrations but with lower mobilities than InSb-like bottom interface in InAs/AlSb heterostructures [227, 253]. Thus, the AlAs interface and alloying were included for the modelling in section 6.3.1. Several solutions were proposed for tackling those issues for InAs growth. An effective As-to-In ratio higher than one can prevent the formation of void defects associated with As etching of the Sb-based layer underneath [254] during the growth of InAs. As a result of the InAs relaxation, mismatch dislocations produce and interfaces become rougher for InAs quantum wells grown on AlSb buffer [255], so the choice of buffer layer is also important. In addition, the channel consists of etched regions on both ends for electrical contact. Due to etch damage and dangling bonds, the surface defect density is typically large

[256] on the etched area. Electron accumulation has been observed after dry etching, which is inherently related to the process-induced structural defects [257].

6.2 Channel Design and Heterojunctions

First and foremost, a general inspection of the entire structure of ULTRA**RAM**™ is presented. Three types of semiconductor heterojunctions organized by band alignment are shown in figure 6.2(a)-(c). Type I of straddling gap, type II of staggered gap and type III with broken gap. In type I, the band gap of the second semiconductor (right) is fully contained in the gap of the first semiconductor (left). For the instance of type II, the gaps from two semiconductors are partially overlapped. For broken band gap, the conduction band from the second semiconductor (right) overlaps with the valence band of the first semiconductor (left), leaving no forbidden energy levels at the interface.

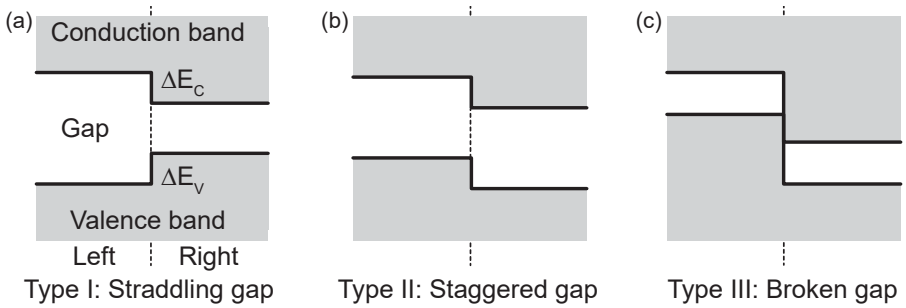

Figure 6.2: Schematic representations of three band alignment types. (a) Straddling gap. (b) Staggered gap. (c) Broken gap. ΔE_C, the conduction band offset; ΔE_V, the valence band offset.

ULTRA**RAM**™ wafers feature multiple epi-layers grown on Si (or GaAs or GaSb) substrates, therefore, it's imperative to conduct a fundamental simulation to get a comprehensive understanding of the layered structure and heterojunctions formed, as device properties depend crucially on the alignment type at the interface [227]. The calculated room temperature band diagram at equilibrium of the top layers of

XPH 2318 using nextnano software is shown in figure 6.3.

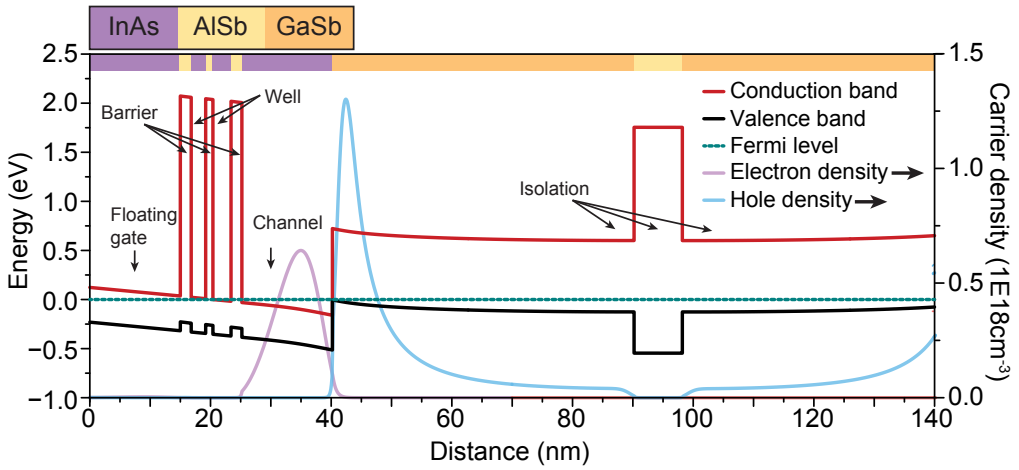

Figure 6.3: Calculated band profile of the epitaxial design of XPH 2318 at 300 K. Corresponding layers are denoted by colour boxes on the top. Electron and hole density are plotted on the right Y-axis.

The primary functioning heterojunction in the structure is InAs/AlSb, from which the TBRT is built up. As seen in figure 6.3, the band lineup for InAs/AlSb heterojunctions is a staggered gap with a large conduction band offset of about 2 eV. It is worth noting that this is for the gamma band. AlSb is indirect band gap, but the gamma band is used here as resonant tunnelling is a coherent process. This lineup of InAs permits a memory operation via resonant tunnelling at very low voltages, empowering the low-switching voltage performance of ULTRA**RAM**™. However, there are shortcomings from the small bandgap of 0.35 eV of InAs and the tiny valence band offset (ΔE_V, around 0.1 eV) at InAs/AlSb interface, with a leakage current via band to band tunnelling (from InAs conduction band to AlSb valence band) reported [258, 259]. This may enable charge transport even when the TBRT is not in a resonant tunnelling condition, thereby reducing the charge-blocking capability of the interface and consequently degrading the memory retention performance. The entire channel region is below Fermi level as indicated by the electron density profile, showing the normally-on property.

Another crucial interface is InAs/GaSb below the channel layer. The band alignment at InAs/GaSb interface is a type II broken gap where resonant interband coupling [260] or band to band tunnelling [261] has been reported, as seen by the hole density peak at the interface on GaSb side. This elucidates the presence of hole current in the GaSb layer in relation to the over-etched channel issue discussed in section 5.3.2, which significantly impacts memory performance, particularly the readout characteristics. It also underscores the necessity of incorporating hole-blocking layers in the vertical direction, thereby explaining the inclusion of isolation units in the wafer design. Further investigation is required into the layer beneath the InAs channel, with a focus on balancing the use of layers preferred for etching monitoring and the electrical performance of the InAs channel.

The last important interface from the layered structure is the GaSb/AlSb heterojunction in the isolation unit which plays a key role in blocking the problematic current vertically through the device introduced by the GaSb layer as mentioned just above. The band alignment at GaSb/AlSb heterojunction is a straddling gap. As shown by the hole density curve, the presence of a hole barrier from the AlSb insertion suppresses the hole density effectively. The GaSb/AlSb interfaces in the design provide effective device-to-device isolation following the mesa etching process.

6.3 Simulations of the Tunnelling Layers in UL-TRARAM™ Devices

In this section, the transport simulation focuses on the TBRT region only which was treated as a resonant tunnelling diode to study the dynamic tunnelling process. The left and right contacts are configured next to the channel and the floating gate layer, respectively. The basic TBRT configuration used in simulation is listed in table 6.1. The InAs in the channel and the floating gate were n-type doped to 1×10^{15} cm^{-3}. Scattering was included in the InAs layers for the transport simulation. Strain was not included in all configurations performed. The default temperature is 300 K

unless stated otherwise. The focus of the work was alloying at the interface, in the barrier and in the quantum well.

Layer	Channel	Barrier 1^{st}	Quantum well 1^{st}	Barrier 2^{nd}	Quantum well 2^{nd}	Barrier 3^{rd}	Floating gate
Material	InAs	AlSb	InAs	AlSb	InAs	AlSb	InAs
Thickness	15	1.8	3.0	1.2	2.4	1.8	15

Table 6.1: Layer details of the TBRT configuration for nextnano simulation.

6.3.1 Interface Alloying

The obtained alloying condition from TEM analysis is in a quaternary form of $Al_xIn_{1-x}As_{1-y}Sb_y$ for the AlSb barrier and the InAs/AlSb interface. Since there is no available database for quaternary materials of the specific form acquired by the compositional analysis, as a workaround, an inserted interface of a 0.6-nm AlAs layer was first carried out for simulation. It was designed to form an InAs/AlAs/AlSb structure to emulate the alloyed InAs/AlSb, but the total thickness of the TBRT was preserved to be the same as the primary design. The consideration is due to the preferential formation of tensile AlAs-type interface in InAs/AlSb structures. Detailed layer configurations used in the simulation of primary design and AlAs insertion are shown in figure 6.4(a)-(b).

(a) Primary

Layer	InAs	AlSb	InAs	AlSb	InAs	AlSb	InAs
Thickness (nm)	15	1.8	3.0	1.2	2.4	1.8	15

(b) AlAs-0.6 nm

Layer	InAs	AlAs	AlSb	AlAs	InAs	AlAs	AlSb	AlAs	InAs	AlAs	AlSb	AlAs	InAs
Thickness (nm)	14.7	0.6	1.2	0.6	2.4	0.6	0.6	0.6	1.8	0.6	1.2	0.6	14.7

Figure 6.4: (a) Primary design and (b) AlAs-0.6 nm layer configuration used for the TBRT simulation.

Figure 6.5(a) pictures the transmission coefficient as a function of energy where a raised resonant level is observed for the case of the AlAs insertion at the InAs/AlSb interface, which can be explained by an effective reduction in the InAs quantum well thicknesses, raising the energies of the confined states [214].

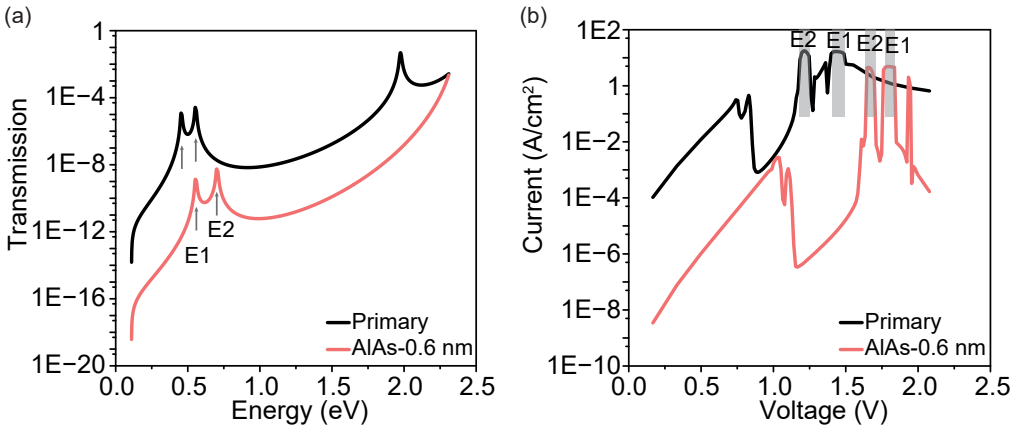

Figure 6.5: (a) Transmission and (b) current-voltage characteristic for the interface alloying with insertion of a 0.6-nm AlAs layer, showing raised resonant levels and reduced transmission coefficient and current intensity. The grey boxes in (b) correspond to the alignments of two resonant levels in (a).

Peaks in transmission curves mark the resonant levels in the TBRT, the alignment to those levels results in different current density peaks seen in the current density plot indicated by grey boxes. In figure 6.5(b), similar shifting to a higher voltage for resonance current density peaks was obtained from the simulation. The gate voltage is 30% higher for the tunnelling with the presence of AlAs interface, raising P/E voltage of the TBRT operation. The overall reduction in current can be attributed to an effective thickening of the barrier due to the insertion of the AlAs.

6.3.2 Barrier Alloying

The incorporation of As into AlSb barrier by replacing Sb was then considered for the barrier alloying due to the higher vulnerability of the ultra-thin thickness and

the very large InAs/AlSb conduction band offset to study its significance on TBRT operation.

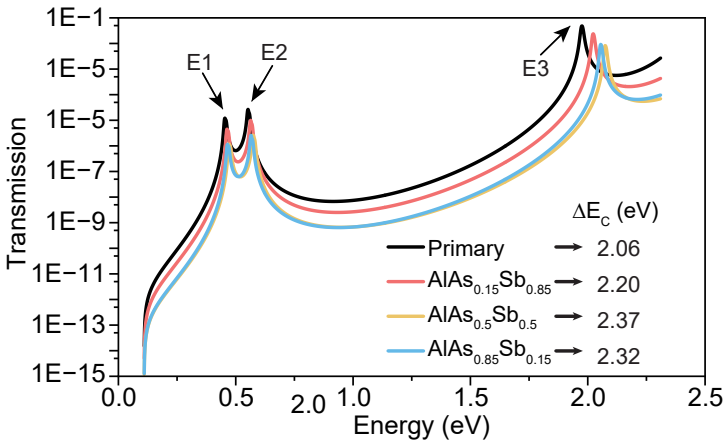

Figure 6.6: Transmission through the TBRT with varying AlSb barrier alloying as a function of energy. The higher As fraction in the barrier pushes resonant levels slightly higher. ΔE_C, the conduction band offset. E1, E2 and E3 are three resonant levels.

Figure 6.6 shows the calculated transmission curves at zero bias for three variations of $AlAs_xSb_{1-x}$. Starting from the primary AlSb with no alloying, the As/Sb ratio is monotonically increased to 0.85/0.15. In comparison to the original AlSb, the incorporation of As/Sb into AlSb barrier has a remarkably limited effect on the transmission coefficient with only a slight reduction for all three As/Sb ratio conditions. Meanwhile, it can be seen that the transmission initially drops with increasing As/Sb ratio, but levels off at high As/Sb ratio, with transmission curve of As/Sb ratio at 0.85/0.15 being very similar to that at 0.5/0.5. This is consistent with the non-monotonic barrier height change with increasing As/Sb ratio as listed in the figure. In particular, the quasi-bound state in the first quantum well state shows negligible change against the As incorporation in terms of resonant energy level. The resonant level for E3 with the highest energy is pushed higher, but since it is not contributing to the tunnelling, such a shift has no impact on the TBRT operation.

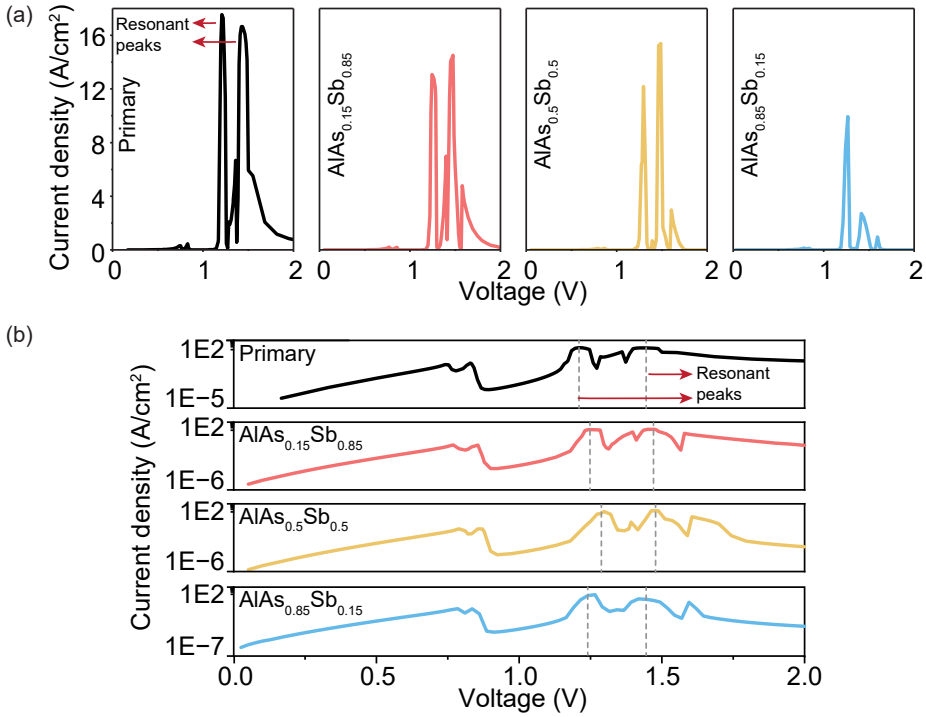

Figure 6.7: (a) Linear and (b) logarithmic plot for current-voltage characteristics of the TBRT with varying AlSb barrier alloying as a function of energy, showing limited changes from As incorporation into AlSb layers.

Figure 6.7(a)-(b) shows the corresponding current-voltage characteristics for all three barrier alloying conditions. As the resonant levels are not changed by As alloying, as seen in the J-V curves, the current density peaks remain generally the same as the primary design, indicating a restricted impact from $AlAs_xSb_{1-x}$ on the TBRT operation, even with an 85% Sb replacement by As in all barriers. A slightly reduced current density is observed compared to the primary design, as a consequence of the diminished transmission coefficient. The elevated As ratio pushes the current peaks to higher voltage levels until As/Sb incorporation of 0.5/0.5, with the shift reverts at high As/Sb ratio of 0.85/0.15, as seen in the transmission changes. Overall As incorporation into the barriers of the TBRT shows a very limited effect on P/E operation of ULTRA**RAM**™.

6.3.3 Quantum Well Alloying

In this section, simulations with two scenarios for the quantum well layers of the TBRT were carried out in the form of $Al_xIn_{1-x}As$ and $InAs_{1-y}Sb_y$ to examine the impact on the memory operation of ULTRA**RAM**™.

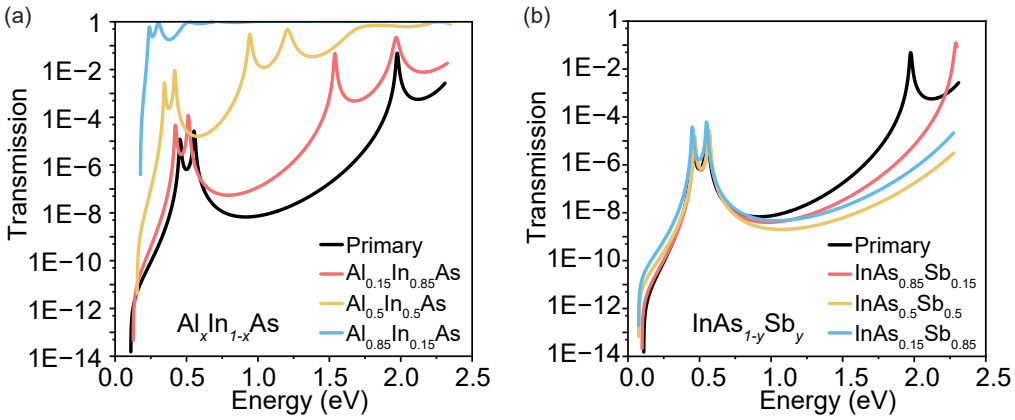

Figure 6.8: Transmission of the TBRT with varying InAs quantum well alloying as a function of energy for (a) $Al_xIn_{1-x}As$ and (b) $InAs_{1-y}Sb_y$. Al fractional composition higher than 0.5 leads to a decaying transmission while Sb alloying into InAs has a limited effect on the first two resonant levels.

Three incremental ratios (0.15/0.85, 0.5/0.5 and 0.85/0.15) were considered for each ternary composition. Figure 6.8(a) shows the transmission plot for $Al_xIn_{1-x}As$. A clear gradual deterioration of the transmission contrast that is characteristic of the TBRT can be observed where the transmission peaks are transitioned to lower energies and higher transmission coefficients are observed with increasing Al incorporation into InAs quantum well. The rising transmission level rapidly reaches 1 at less than 0.5 eV, highlighting a poor charge blocking capability of the TBRT, which is destructive to its effective operation.

The Al alloying condition reduces the energy for tunnelling condition and also increases the transmission notably. This is consistent with what is seen in figure 6.9(a)-(b), where the current density increases by about two orders of magnitude for the 0.85 Al-fraction case. The Al incorporation into InAs reduces the band

offset dramatically, resulting in a tiny band offset of only 0.34 eV at Al/In ratio of 0.85/0.15. For the instance of Al/In ratio of 0.85/0.15, the peak current density reaches up to 7000 kA/cm^2 and barely shows any NDR effect from resonances with quantum wells. The vanishing peaks and valleys from the J-V characteristics demonstrate the damaging effect of the incorporation of Al into InAs well for the TBRT operation.

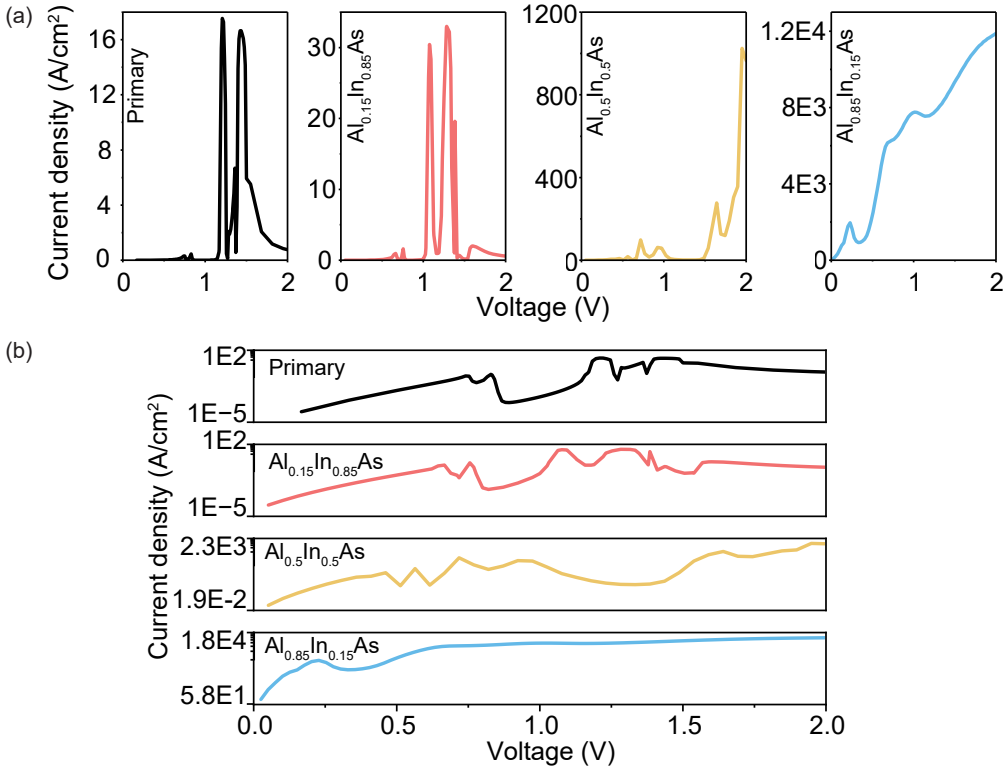

Figure 6.9: (a) Linear and (b) logarithmic plot for current-voltage characteristics of the TBRT with varying InAs quantum well alloying as a function of energy for $Al_xIn_{1-x}As$. The degradation of the resonant peaks suggests the disappearance of the charge blocking capability of the TBRT.

Figure 6.8(b) depicts Sb alloying into InAs quantum well, showing the transmission as a function of energy. Different to the situation of Al, Sb incorporation into InAs shows limited impact on TBRT operation regarding the transmission, with negligible changes with respect to both transmission coefficient and resonant energy.

The E3 resonant level is again pushed to a higher level, however, it's not relevant to memory operation.

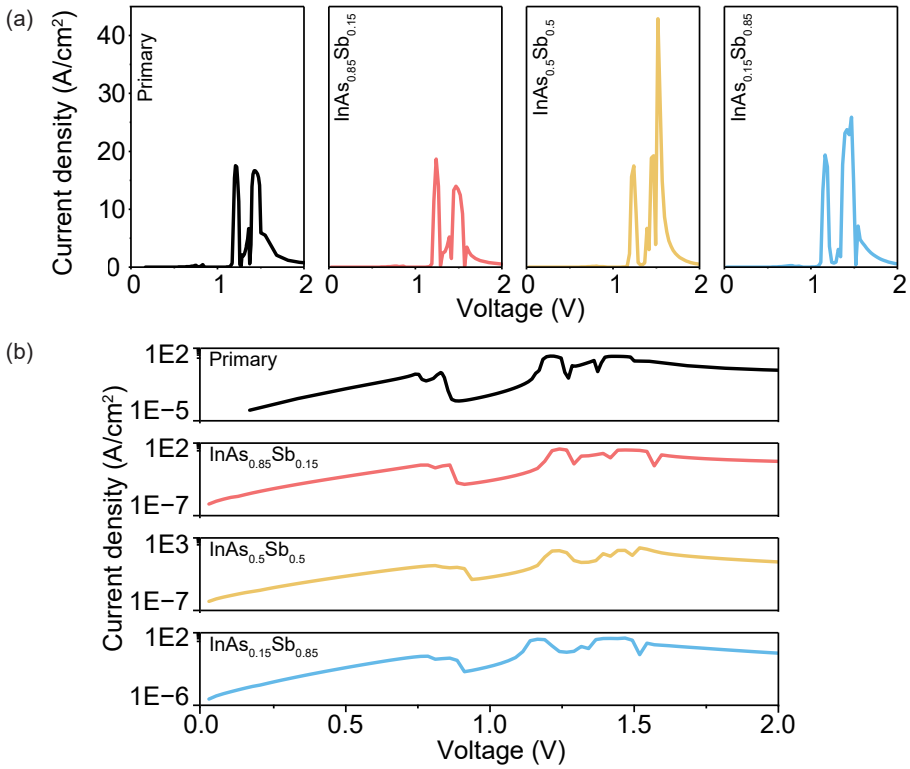

Figure 6.10: (a) Linear and (b) logarithmic plot for current-voltage characteristics of the TBRT with varying InAs quantum well alloying as a function of energy for $InAs_{1-y}Sb_y$. No significant impact from Sb incorporation is observed.

The corresponding current-voltage characteristics for $InAs_{1-y}Sb_y$ are shown in figure 6.10(a)-(b). The voltages for current density peaks of all three variations of Sb ratios are lined up to primary design, showing a high degree of consistency with the primitive binary InAs, except for a slight change in the current density magnitude. The Sb replacement of As exerts a minimal influence on the conduction band edge, such that the band offset conditions for InAs and $InAs_{1-y}Sb_y$ remain almost the same.

The actual scenario, as revealed by TEM observations, corresponds to a quaternary configuration involving alloying within the AlSb barrier. This condition

resembles a combination of barrier alloying and the presence of an AlAs interface. Both the AlAs interface and barrier alloying scenarios suggest that a higher operating voltage is necessary for effective memory function. This requirement may partially account for the limited memory window observed, as the conventional operating voltage of 2.6 V used in all the memory characterisations may be insufficient for optimal TBRT performance. Nevertheless, none of these cases undermine the concept of the TBRT operation. It is worth noting that, to ensure consistent device performance, interface engineering—targeted at achieving sharper interfaces and minimising intermixing—may prove advantageous in future investigations. Furthermore, as previously discussed, the degree of asymmetry may be further optimised to enable lower switching voltages; however, care must be taken to avoid the occurrence of resonance under zero-bias conditions.

6.4 Summary

In conclusion, different scenarios of alloying in the interface, barrier and quantum well of the TBRT were simulated to study its effect on ULTRA**RAM**™ memory. For interface alloying simulated by the inclusion of an AlAs layer, the resonant energy levels are raised around 30%, so an elevated voltage would be required for P/E operation through the TBRT. With respect to the As incorporation into the barrier and Sb incorporation into the quantum well, a similar effect of slightly raising the resonant energy levels was observed, playing a limited role in TBRT operation. On the other hand, Al incorporation in InAs can be highly destructive for the TBRT in the situation of higher ratios. However, fortunately, this case was not seen in the cross-sectional TEM analysis, and is thus considered to be unlikely, at least for MBE-grown material.

Due to the limitation of database available in nextnano, other ternary and quaternary compositions were not simulated. No deleterious effect was found in all simulated alloying conditions, except for the Al incorporation into InAs which

was not supported by the TEM analysis. This further confirms the robustness of memory operation and the charge blocking capability of the TBRT structure, and hence, ULTRA**RAM**™'s resilience against growth fluctuation and errors. This simulation addresses the gap in the previous simulation study concerning layer variation, thereby enhancing the overall comprehensiveness of the simulation work of ULTRA**RAM**™. Furthermore, the simulation, informed by the TEM observations, serves as a valuable reference for comparison with the memory measurement results. These findings suggest that higher voltages and asymmetric program/erase conditions may be required for more comprehensive characterisation of the memory behaviour, and they provide useful guidance for future wafer growth and process optimisation.

Chapter 7

Conclusions and Future Work

7.1 Conclusions

Over the years, significant progress has been made on ULTRA**RAM**™ through both experimental and simulation efforts, with the aim of enabling commercialisation and practical application. Nevertheless, ongoing developments continue to advance ULTRA**RAM**™ technology, with sustained efforts in both domains directed towards improving scalability and enhancing memory performance. Future work will focus on further optimisation, increased scalability, and the progression towards array-level design.

In this work, fabrication, characterisation and simulation have been carried out to develop a scalable fabrication design and achieve enhanced performance for ULTRA**RAM**™ memory.

Three progressive designs have been discussed and analysed. Starting from the legacy design using wet-etch, the introduction of all-dry etching through to the self-aligned design marks the ability of a scalable fabrication of ULTRA**RAM**™. Directed by the proposed explanation of the resistive-gap, a preliminary fabrication using an improved compact design has shown promising results regarding the two challenges remaining in the previous designs, the scalability and the resistive-gap, with the feasibility to further scale down the device to 5 μm gate size and the size

of the resistant gap reduced by a notable factor of 10. Specifically, over-etching of the 10-nm channel, which outweighs all other conceivable determinants within the memory fabrication with regard to the resistive-gap issue, has been reduced to 57% throughout the optimisation, by a refined recipe and fine-tuned epitaxial design. Modelling and experimental validation of end-point detection with shorter wavelength of 405 nm have shown further enhancement to the pressing issue of the channel over-etching, where the TBRT layers could be resolved during the ICP etching, nearly eliminating the over-etching issue and securing a thicker channel for the memory readout performance.

Multiple approaches including electrical measurement and X-ray nanoprobe technique have been used to characterise the as-fabricated memories. The mA scale control gate leakage issue has been solved by the adoption of a Si_3N_4 hard mask, as supported by systemic electrical measurements. Endurance and retention characterisation of ULTRA**RAM**™ memories from three batches of fabrication validate the self-aligned design with the best endurance memory window of ~ 10 μA and shortest operation pulse of 5 ms. The 10 times larger endurance memory window, observed in the comparison between the XPH 1823 and XPH 2213 batches, provides compelling evidence for the effectiveness of the processing optimisations implemented in etching control and fabrication design. In addition, an X-ray nanoprobe technique has been employed to investigate mechanisms underlying the cycling-failure of ULTRA**RAM**™ memory, showing that no correlation was found between In-related defects and the memory cycling.

Simulations have been performed concerning the TBRT interface disturbance observed from the TEM characterisation, in the form of interfacial insertion, as well as quantum well and barrier alloying, with alloy compositions in the latter two ranging from 15% to 85%. Though Al incorporation exceeding 50% into InAs can cause a detrimental effect to TBRT operation, such a scenario is not supported by TEM analysis. The calculated transmission coefficients and current characteristics of AlAs interfacial layer, $InAs_xSb_{1-x}$ and $AlAs_xSb_{1-x}$ suggest a slight

shift of the tunnelling voltage-an increase of approximately 30%, more than 13%, and less than 6%, respectively-implying a higher memory program/erase voltage for ULTRA**RAM**™ but no sign of undermining the TBRT operation, thereby affirming its robustness.

The demonstrated scalability of the fabrication design, the enhanced performance observed from the characterisation and the reliability evidenced in the simulation, collectively support advancements in the development of ULTRA**RAM**™. The findings in this work pave the way for further nanometre scaling with high performance and allows ULTRA**RAM**™ to unleash its full potential, an essential step for ULTRA**RAM**™ towards commercialisation.

7.2 Future Work

This work presents promising results for the development of ULTRA**RAM**™, however, challenges remain in the nanometre scaling and commercialisation.

While the proposed design has demonstrated the potential to tackle the resistive-gap issue, the full implementation of proposed compact design requires further work and detailed optimisation, especially on the sidewall profile of the gate stack that is crucial to a successful building of the wrap-around contact. In addition, despite the advantage of the end-point detection with short wavelengths over the channel etching, given that all the dry etchings will be finished within the ICP chamber, the compatibility of the short wavelengths with other etching materials, such as alumina and PMMA, require further verification. Concerning the wafer layout, further work is required to identify an optimal virtual-substrate layer beneath the InAs channel that ensures superior InAs performance while maintaining satisfactory accuracy in etch monitoring, or to explore optimised doping strategies, such as the use of remote donor layers, for achieving optimal channel conductivity.

With regard to the simulation, given that nextnano MSB runs on a single band model, it is worthy to perform the simulation with a multi-band model to further

investigate TBRT operation, as it was reported that it's better to include the valence band to get an accurate calculation of the tunnelling properties [249]. Moreover, as confirmed by the TEM compositional analysis, the AlSb barrier is a quaternary form, together with the interface roughness, the two factors can be taken into account in simulation to gain further insight of the impact to the TBRT from the interface disturbance. In addition, further investigation into the optimal number of the barriers, as well as the device performance at elevated temperatures, may prove beneficial in fully optimising the tunnelling structure.

For the long-term and the final goal of nanometre ULTRA**RAM**™ with normally-off channel and array design, it is necessary-alongside the exploration of alternative III–V channel materials such as InGaAs-to incorporate EBL into the fabrication process, and this may require extreme precise control of the inevitable channel etching which is feasible with atomic layer etching. While technical challenges persist, the integration of EBL, the compact device design, and enhanced etching control is anticipated to enable high-performance ULTRA**RAM**™, with switching time below 1 ns and switching energy under 10 aJ, at gate dimension less than 20 nm.

Appendix A

Fabrication Details

A.1 Epitaxial Designs

The epitaxial design and the thickness of each layer measured by TEM of XPH 1452, XPH 1823, XPH 1896, XPH 2093, XPH 2213 and XPH 2318 are listed in table A.1, table A.2, table A.3, table A.4 and table A.5, and table A.6, respectively.

Layer	Material	Nominal thickness (nm)	Thickness measured by TEM (nm)
Floating gate	InAs	10	9.7
Quantum barrier	AlSb	1.8	2.5
Quantum well	InAs	2.4	1.8
Quantum barrier	AlSb	1.2	1.8
Quantum well	InAs	3.0	2.5
Quantum barrier	AlSb	1.8	2.5
Channel	InAs (n-type 5×10^{18})	10	10.7
	GaSb	20	19.6
	AlSb	8	7.2
Back gate	InAs (n-type)	50	58.7
2-step buffer	GaSb (hot)	540	-
	GaSb (cold)	1400	-
Nucleation	AlSb	5.2	-
Substrate	Si	$\sim$ 380 μm	-

Table A.1: Detailed layout of XPH 1452 wafer for ULTRA**RAM**™, utilised in the wet-etching fabrication.

Layer	Material	Nominal thickness (nm)	Thickness measured by TEM (nm)
Floating gate	InAs	10	8
Quantum barrier	AlSb	1.8	2.2
Quantum well	InAs	2.4	3.3
Quantum barrier	AlSb	1.2	1.6
Quantum well	InAs	3.0	2.
Quantum barrier	AlSb	1.8	2.2
Channel	InAs (n-type 5×10^{18})	10	9.2
	GaSb	20	-
Isolation × 4	AlSb	8	-
	GaSb	50	-
	$Al_{0.7}Ga_{0.3}Sb$	30	-
	GaSb	50	-
2-step buffer	GaSb (hot)	540	-
	GaSb (cold)	1400	-
Nucleation	AlSb	5.2	-
Substrate	Si	$\sim$ 380 μm	-

Table A.2: Detailed layout of XPH 1823 wafer for ULTRA**RAM**™, featuring the introduction of isolation layers.

Layer	Material	Nominal thickness (nm)	Thickness measured by TEM (nm)
Floating gate	InAs	10	9.7
Quantum barrier	AlSb	1.8	1.9
Quantum well	InAs	2.4	2.4
Quantum barrier	AlSb	1.2	1.7
Quantum well	InAs	3.0	3.1
Quantum barrier	AlSb	1.8	2.7
Channel	InAs (n-type 5×10^{18})	10	10
Isolation × 3	AlSb	8	8.3
	GaSb	50	49.3
	$Al_{0.7}Ga_{0.3}Sb$	30	35
	GaSb	50	50
2-step buffer	GaSb (hot)	540	-
	GaSb (cold)	1400	-
Nucleation	AlSb	5.2	-
Substrate	Si	$\sim$ 380 μm	-

Table A.3: Detailed layout of XPH 1896 wafer for ULTRA**RAM**™, showing the typical design of the floating gate, TBRT, channel, and isolation layers.

Layer	Material	Nominal thickness (nm)	Thickness measured by TEM (nm)
Floating gate	InAs	10	10.2
Quantum barrier	AlSb	1.8	2.2
Quantum well	InAs	2.4	2.7
Quantum barrier	AlSb	1.2	1.6
Quantum well	InAs	3.0	3.3
Quantum barrier	AlSb	1.8	2.3
Channel	InAs (n-type 5×10^{18})	15	17.6
For reflectance	AlSb	8	8.7
	GaSb	20	20.2
Isolation	AlSb	5	5.9
	GaSb	50	52.4
	$Al_{0.7}Ga_{0.3}Sb$	30	32
	GaSb	50	53
Isolation × 2	AlSb	8	8.2
	GaSb	50	-
	$Al_{0.7}Ga_{0.3}Sb$	30	-
	GaSb	50	-
2-step buffer	GaSb (hot)	540	-
	GaSb (cold)	1400	-
Nucleation	AlSb	5.2	-
Substrate	Si	$\sim$ 380 μm	-

Table A.4: Detailed layout of XPH 2093 wafer for ULTRA**RAM**™. The additional layers between the channel and the isolation unit are introduced to improve signal identification during ICP etching.

Layer	Material	Nominal thickness (nm)	Thickness measured by TEM (nm)
Floating gate	InAs	15	16
Quantum barrier	AlSb	1.8	2.254
Quantum well	InAs	2.4	2.45
Quantum barrier	AlSb	1.2	1.568
Quantum well	InAs	3.0	2.94
Quantum barrier	AlSb	1.8	2.254
Channel	InAs (n-type 5×10^{18})	15	16.855
	GaSb	20	23.905
Isolation × 4	AlSb	8	8.17
	GaSb	50	64.25
	$Al_{0.7}Ga_{0.3}Sb$	30	30.76
	GaSb	50	64.52
2-step buffer	GaSb (hot)	540	-
	GaSb (cold)	1400	-
Nucleation	AlSb	5.2	-
Substrate	Si	$\sim$ 380 μm	-

Table A.5: Detailed layout of XPH 2213 wafer for ULTRA**RAM**™, designed with increased channel thickness.

Layer	Material	Nominal thickness (nm)	Thickness measured by TEM (nm)
Floating gate	InAs	15	13.061
Quantum barrier	AlSb	1.8	2.23
Quantum well	InAs	2.4	2.336
Quantum barrier	AlSb	1.2	1.805
Quantum well	InAs	3.0	2.442
Quantum barrier	AlSb	1.8	2.23
Channel	InAs	15	15.398
	GaSb	20	19.991
Isolation × 4	AlSb	8	8.178
	GaSb	50	-
	$Al_{0.7}Ga_{0.3}Sb$	30	-
	GaSb	50	-
2-step buffer	GaSb (hot)	540	-
	GaSb (cold)	1400	-
Nucleation	AlSb	5.2	-
Substrate	Si	$\sim$ 380 μm	-

Table A.6: Detailed layout of XPH 2318 wafer for ULTRA**RAM**™, incorporating undoped InAs channel.

A.2 Etching Recipes

Recipes used for ICP etching are listed in table A.7. A typical etching process contains four steps including pump, gas stabilisation, etch and pump. Only parameters for the etch step is listed in the table. The recipe for the hardened S1813 etching is the same as the PMMA etching. The mesa etchings were done with same gases as in the channel etching but with higher power. Recipes used for RIE etching are listed in table A.8. Note that etching time may vary from batch to batch, especially for O_2 plasma cleaning, depending on the sample. A typical RIE etching process consists of multiple stages including pumping, stablisation, striking, etching, purging and pump. Parameters listed in the table are for the etching step. Solutions used for wet-etchings are listed in table A.9.

Parameters	Channel CH_4-based	Optimisation CH_4-based	Channel BCl_3-based	Al_2O_3	PMMA
Pressure (mTorr)	10	10	6	6	20
RF power (W)	100	100	25	50	100
ICP power (W)	150	150	110	110	200
Table temp (°C)	10	10	10	10	10
BCl_3 (sccm)	-	-	10	10	-
CH_4 (sccm)	5	12	-	-	-
H_2 (sccm)	30	30	-	-	-
Ar (sccm)	5	6	10	10	-
O_2 (sccm)	-	-	-	-	45
Etching end-point	Refer to corresponding simulations				

Table A.7: ICP recipes for etching various materials used in the ULTRA**RAM**™ fabrication.

Parameters	O_2 Plasma	Si_3N_4	Al_2O_3	Ta	PMMA
Pressure (mTorr)	100	55	50	100	20
HF power (W)	150	150	200	150	70
Table temp (°C)	25	25	25	25	25
CHF_3 (sccm)	-	50	-	-	-
CF_4 (sccm)	-	-	30	30	-
O_2 (sccm)	40	5	5	5	45
Etching time (min)	2	3.5	3	1.75	3

Table A.8: RIE recipes for various materials used in the ULTRA**RAM**™ fabrication. The etching time for Si_3N_4, Al_2O_3, Ta and PMMA is based the nominal thickness of 180 nm, 15 nm, 63 nm and 200 nm, respectively.

Material	Solution
InAs	Citric acid : H_2O_2 : H_2O = 1 : 3 : 1 (volume ratio)
AlSb	MF-319

Table A.9: Solutions used for wet-etching of InAs and AlSb.

A.3 Lithography Details

Spin-coating parameters used for optical mask lithography and LW are listed in table A.10. Developing procedures for each combination of photoresists are listed in table A.11. Depending on the stock, MF-319 and MF-CD26 are used interchangeably. Lithography patterns used for the improved design with a gate size of 5 μm × 5 μm are shown in figure A.1.

Photoresist	S1813	LOR 3A	PMMA
Spinning time (s)	60	30	60
Spinning acc (rpm/s)	3000	1500	1000
Spinning speed (rpm)	6000	3000	2000
Baking temp (°C)	115	170	180
Baking time (min)	2	5	2
Developer	MF-CD26	MF-CD26	H_2O/IPA = 3:1
Stripping	Acetone	Remover 1165	Acetone

Table A.10: Spinning parameters for photoresists used in this work.

Photoresist	S1813	S1813 + S1813	S1813 + LOR 3A
Lithography	Optical mask lithography or LW		
Developer time	1min	1min	1min
H_2O rinse	1min	1min	1min
Post-exposure bake	No	No	5 mins at 125 °C
Developer time	No	No	1min
H_2O rinse	No	No	1min
N_2 drying	Yes	Yes	Yes

Table A.11: Developing procedures for different photoresist combinations.

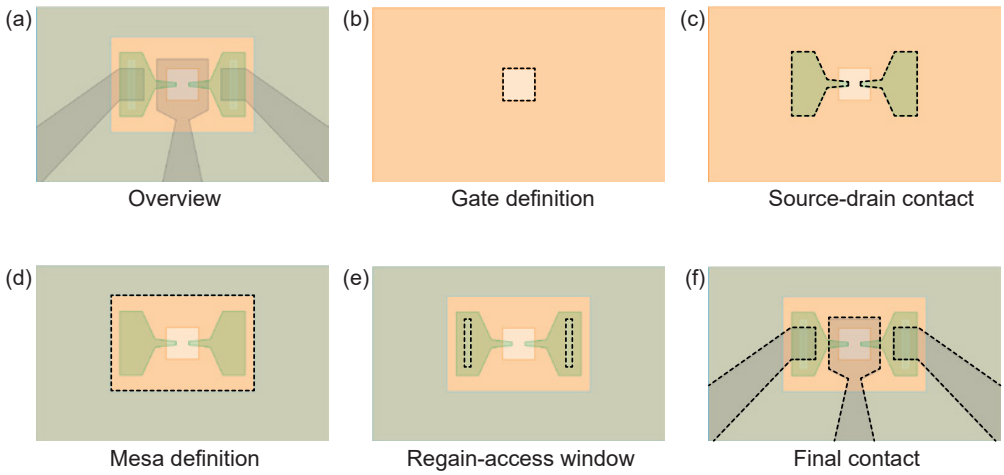

Figure A.1: Lithography patterns for the improved compact design. (a) Overview of the device design. (b) Gate definition. The region outside the gate (the dashed box) is to be etched. (c) Source-drain contact. The finger-shaped contact is patterned for the metallisation. The overlap between the gate stack and the contact ensures a close contact to the gate after the lift-off process. (d) Mesa definition. The region outside the mesa (the dashed box) is to be etched. (e) Regain-access window. Two windows are opened on the source-drain contact by removing the alumina passivation. (f) Final contact. Final metallisation is done to bridge the source-drain contact to contact pads for subsequent measurements.

A.4 Dielectric Recipes

Key parameters used for PECVD deposition of Si_3N_4 are listed in table A.12. Two recipes used for alumina deposition by ALD are listed in table A.13.

Parameters	Pump	Stablisation	Strike	Pressure	Pressure	Deposit	Pump
Pressure (mTorr)	0	25	25	17	10	10	0
HF power (W)	0	0	50	0	0	0	0
ICP power (W)	0	0	250	250	250	250	0
Table temp (°C)	140	140	140	140	140	140	25
SiH_4 (sccm)	0	5.3	5.3	5.3	5.3	5.3	0
N_2 (sccm)	0	4.5	4.5	4.5	4.5	4.5	0
Time (min)	60	1	0.25	0.25	0.25	3	2

Table A.12: PECVD recipe of each step for Si_3N_4 deposition used in the ULTRA**RAM**™ fabrication.

Recipes	Heaters temp (°C)	H_2O Pulse (s)	Wait (s)	TMA pulse (s)	Wait (s)	Repetition (cycles)	Carrier N_2 (sccm)
80 °C	80	0.015	10	0.015	10	200	20
150 °C	150	0.015	5	0.015	5	150	20

Table A.13: ALD deposition recipes for Al_2O_3 used in the ULTRA**RAM**™ fabrication.

Appendix B

Basics of Simulation

B.1 Etching Simulation

The optical constants, including refractive index n and extinction coefficient k, of materials involved in the simulated structures are listed in table B.1 and table B.2. The numbers for $Al_{0.7}Ga_{0.3}Sb$ are taken from a similar composition, $Al_{0.5}Ga_{0.5}Sb$. All 206 nm data are taken from 206.6 nm wavelength. For PMMA and alumina used in the simulation at 670 nm wavelength, the refractive indices are 1.4995 and 1.7643 [262], respectively.

An etch-through test prior to device processing is paramount for different structures to circumvent the possible deviation between the simulation curve and the experimental result due to various reasons. As a general etching guide, the comparison between the reflectance from an etch-through of PMMA/XPH 2318 structure and the simulation is plotted in figure B.1. The slight difference can be explained by the thickness variation between the simulation and the actual thickness of the PMMA used.

	InAs		AlSb		GaSb	
Parameters	n	k	n	k	n	k
206 nm	1.4340	2.1120	1.0648	2.2263	0.85237	2.3048
340 nm	3.0485	1.7218	3.9363	2.8857	3.7617	2.7714
365 nm	3.0044	1.7913	3.9746	2.6414	3.7944	2.4270
405 nm	3.1419	1.9863	4.5510	2.0629	3.7531	2.1245
488 nm	4.2214	1.8272	4.6932	0.53068	4.1756	2.2829
633 nm	3.9637	0.60909	3.8638	3.5982e-3	5.1639	1.1574
670 nm	3.8852	0.55175	3.7714	2.5179e-3	4.9387	0.72390
905 nm	3.5466	0.21482	3.4557	7.0993e-2	4.0554	0.30682

Table B.1: Optical parameters of InAs, AlSb and GaSb used in the etching simulation with various wavelengths [262].

	$Al_{0.5}Ga_{0.5}Sb$		Si	
Parameters	n	k	n	k
206 nm	1.2460	2.3050	1.0100	2.9090
340 nm	3.9237	2.9102	5.2298	3.0206
365 nm	4.0740	2.4490	6.5271	2.6672
405 nm	4.0072	2.1570	5.4376	0.34209
488 nm	4.5121	1.6517	4.3707	8.0068e-2
633 nm	4.6235	0.22167	4.3707	8.0068e-2
670 nm	4.3872	0.12808	3.8224	1.4554e-2
905 nm	3.8623	2.8824e-3	3.8823	1.9589e-2

Table B.2: Optical parameters of $Al_{0.5}Ga_{0.5}Sb$ and Si used in the etching simulation with various wavelengths [262].

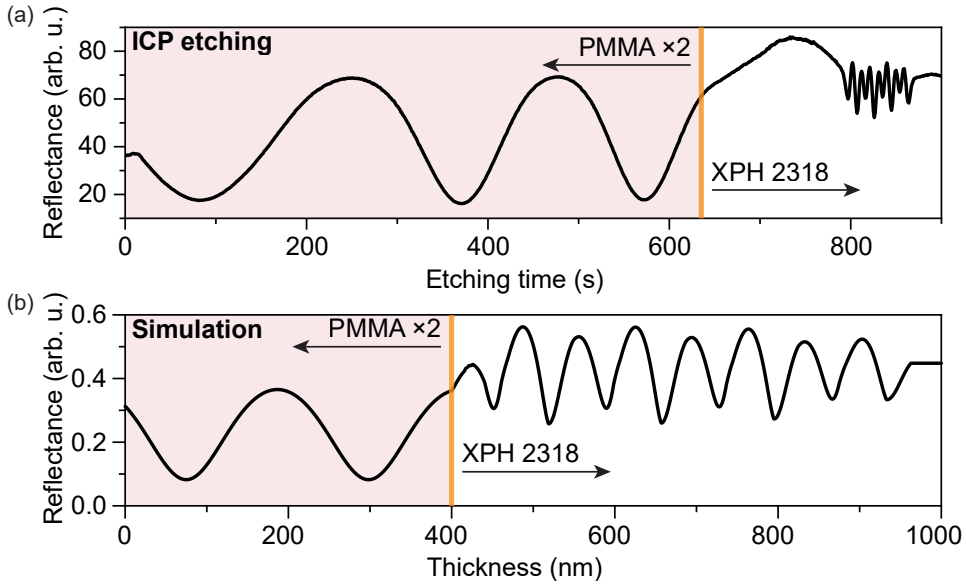

Figure B.1: (a) Etching reflectance from the etch-through of PMMA/XPH 2318 structure. (b) The simulation for the structure in (a). Reflectance fingerprints of the PMMA are superimposed on the beginning of the known wafer reflectance. The orange line marks the stop line for end-point detection. Double layers of PMMA were used in the etching with a nominal thickness of 400 nm, which was used for the simulation in (b). The deviation between (a) and (b) can be attributed to the variation of PMMA thickness which depends on the spinning and baking conditions.

B.2 nextnano Simulation

B.2.1 Material Parameters

Material parameters used in the simulation are listed in table B.3.

B.2.2 Interpolation of Ternary Compounds

Properties of zinc blende type ternary alloys simulated in nextnano, including $AlAs_xSb_{1-x}$, $Al_xIn_{1-x}As$, $InAs_{1-y}Sb_y$, are interpolated by two binary alloys (InAs, InSb, AlSb and AlAs) with bowing parameters. This is considered to be two-component alloy by nextnano, which is constant bowing (quadratic). For a two-

component alloy with the form of AB_xC_{1-x}, the general quadratic form is

$$P_{ABC}(x) = x \cdot P_{AB} + (1-x) \cdot P_{AC} - x(1-x) \cdot b_{ABC}\,, \tag{B.1}$$

where the $P_{ABC}(x)$ is the interpolated property of AB_xC_{1-x}, P_{AB} and P_{AC} are corresponding properties from the two pure components, and b_{ABC} is the bowing parameter for the alloy. The corresponding parameters are listed in table B.4

Parameters	InAs	AlSb	AlAs	InSb
CB offset (eV)	1.937	3.996	4.049	2.255
VB offset (eV)	1.39	1.385	0.857	1.750
Band gap (eV)	0.417	2.386	3.099	0.235
Effective mass (m_0)	0.026	0.14	0.15	0.0135
Material density (kg/m^3)	5.61e3	4.26e3	3.72e3	5.77e3
Deformation potential (eV)	-6.66	-8.12	-7.4	-6.04
Static dielectric constant	15.15	12.04	10.064	17.5
Optic dielectric constant	12.25	10.24	8.162	15.68
LO phonon energy (eV)	30e-3	42e-3	50e-3	22e-3
LO phonon width (eV)	3e-3	3e-3	3e-3	3e-3

Table B.3: nextnano material parameters for the simulation.

Parameters	$AlAs_xSb_{1-x}$	$Al_xIn_{1-x}As$	$InAs_{1-y}Sb_y$
Components	Al_xAs & $AlSb_{1-x}$	$AlAs_x$ & $In_{1-x}As$	$InSb_x$ & $InAs_{1-x}$
Bow-band-gaps	0.8	0.7	0.67
Bow-conduction-band-energies	-0.91	0.06	0.67
Bow-valence-band-energies	0.417	-0.69	-0.4

Table B.4: Bowing parameters for the simulation.

References

[1] Belinda Dube Simone Bertolazzi. *Spotlight on DRAM*. https://www.yolegroup.com/strategy-insights/spotlight-on-dram/. [Accessed 02-01-2025].

[2] Shimeng Yu et al. "Compute-in-memory chips for deep learning: Recent trends and prospects". In: *IEEE circuits and systems magazine* 21.3 (2021), pp. 31–56.

[3] Jeongdong Choe. "Memory technology 2021: Trends & challenges". In: *2021 international conference on simulation of semiconductor processes and devices (SISPAD)*. IEEE. 2021, pp. 111–115.

[4] H-S Philip Wong and Sayeef Salahuddin. "Memory leads the way to better computing". In: *Nature nanotechnology* 10.3 (2015), pp. 191–194.

[5] Ofogh Tizno et al. "Room-temperature operation of low-voltage, non-volatile, compound-semiconductor memory cells". In: *Scientific reports* 9.1 (2019), p. 8950.

[6] Dominic Lane and Manus Hayne. "Simulations of ultralow-power nonvolatile cells for random-access memory". In: *IEEE Transactions on Electron Devices* 67.2 (2020), pp. 474–480.

[7] Dominic Lane. "ULTRARAM™: Design, Modelling, Fabrication and Testing of Ultra-low-power III-V Memory Devices and Arrays". English. PhD thesis. Lancaster University, 2021. DOI: 10.17635/lancaster/thesis/1465.

[8] D Lane et al. "ULTRARAM: toward the development of a III–V semiconductor, nonvolatile, random access memory". In: *IEEE Transactions on Electron Devices* 68.5 (2021), pp. 2271–2274.

[9] Peter D Hodgson et al. "ULTRARAM: A Low-Energy, High-Endurance, Compound-Semiconductor Memory on Silicon". In: *Advanced Electronic Materials* 8.4 (2022), p. 2101103.

[10] Kahng Dawon. *Electric field controlled semiconductor device*. US Patent 3,102,230. 1963.

[11] Dawon Kahng. "A historical perspective on the development of MOS transistors and related devices". In: *IEEE Transactions on Electron Devices* 23.7 (1976), pp. 655–657.

[12] Isabelle Ferain, Cynthia A Colinge, and Jean-Pierre Colinge. "Multigate transistors as the future of classical metal–oxide–semiconductor field-effect transistors". In: *Nature* 479.7373 (2011), pp. 310–316.

[13] Gordon E Moore. "Cramming more components onto integrated circuits". In: *Proceedings of the IEEE* 86.1 (1998), pp. 82–85.

[14] Dawon Kahng and Simon M Sze. "A floating gate and its application to memory devices". In: *The Bell System Technical Journal* 46.6 (1967), pp. 1288–1295.

[15] Flash Memories. *P. Cappelletti, C. Golla, P. Olivo, and E. Zanoni, Eds.* 1999.

[16] Eli Harari. "The Non-Volatile memory industry-a personal journey". In: *2011 3rd IEEE International Memory Workshop (IMW)*. IEEE. 2011, pp. 1–4.

[17] N Goel et al. "Erase and retention improvements in charge trap flash through engineered charge storage layer". In: *IEEE electron device letters* 30.3 (2009), pp. 216–218.

[18] Jaehoon Jang et al. "Vertical cell array using TCAT (Terabit Cell Array Transistor) technology for ultra high density NAND flash memory". In: *2009 symposium on VLSI technology*. IEEE. 2009, pp. 192–193.

[19] Seung Soo Kim et al. "Review of semiconductor flash memory devices for material and process issues". In: *Advanced Materials* 35.43 (2023), p. 2200659.

[20] Frank R Libsch and Marvin H White. "Charge transport and storage of low programming voltage SONOS/MONOS memory devices". In: *Solid-state electronics* 33.1 (1990), pp. 105–126.

[21] Frank M Wanlass and C T Sah. "Nanowatt logic using field-effect metal-oxide semiconductor triodes". In: *Semiconductor devices: pioneering papers*. World Scientific, 1991, pp. 637–638.

[22] John Von Neumann. "First Draft of a Report on the EDVAC". In: *IEEE Annals of the History of Computing* 15.4 (1993), pp. 27–75.

[23] Chow Wen Tsing and William H Henrich. *Storage matrix*. US Patent 3,028,659. Apr. 1962.

[24] Dov Frohman-Bentchkowsky. "FAMOS—A new semiconductor charge storage device". In: *Solid-State Electronics* 17.6 (1974), pp. 517–529.

[25] Fujio Masuoka. "Technology trend of flash-EEPROM-Can flash-EEPROM overcome DRAM?" In: *1992 Symposium on VLSI Technology Digest of Technical Papers*. IEEE. 1992, pp. 6–9.

[26] Eliyahou Harari. *Electrically erasable non-volatile semiconductor memory*. US Patent 4,115,914. Sept. 1978.

[27] Fujio Masuoka and Hisakazu Iizuka. *Semiconductor memory device and method for manufacturing the same*. US Patent 4,531,203. July 1985.

[28] Anisha Ramesh, Si-Young Park, and Paul R Berger. "90 nm 32 × 32 bit Tunneling SRAM Memory Array With 0.5 ns Write Access Time, 1 ns Read Access Time and 0.5 V Operation". In: *IEEE Transactions on Circuits and Systems I: Regular Papers* 58.10 (2011), pp. 2432–2445.

[29] Young Hoon Son et al. "Reducing memory access latency with asymmetric DRAM bank organizations". In: *Proceedings of the 40th annual international symposium on computer architecture*. 2013, pp. 380–391.

[30] Wooseong Cheong et al. "A flash memory controller for 15μs ultra-low-latency SSD using high-speed 3D NAND flash with 3μs read time". In: *2018 IEEE International Solid-State Circuits Conference-(ISSCC)*. IEEE. 2018, pp. 338–340.

[31] David A Thompson and John S Best. "The future of magnetic data storage techology". In: *IBM Journal of Research and Development* 44.3 (2000), pp. 311–322.

[32] David A Laws. "A company of legend: The legacy of fairchild semiconductor". In: *IEEE Annals of the History of Computing* 32.1 (2010), pp. 60–74.

[33] Zheng Guo et al. "FinFET-based SRAM design". In: *Proceedings of the 2005 international symposium on Low power electronics and design*. 2005, pp. 2–7.

[34] Yoshinobu NAKAGOME and Kiyoo ITOH. "Reviews and prospects of DRAM technology". In: *IEICE TRANSACTIONS on Electronics* 74.4 (1991), pp. 799–811.

[35] Alessio Spessot and Hyungrock Oh. "1T-1C dynamic random access memory status, challenges, and prospects". In: *IEEE Transactions on Electron Devices* 67.4 (2020), pp. 1382–1393.

[36] Aniruddha N Udipi et al. "Rethinking DRAM design and organization for energy-constrained multi-cores". In: *Proceedings of the 37th annual international symposium on Computer architecture*. 2010, pp. 175–186.

[37] Fujio Masuoka et al. “A new flash E 2 PROM cell using triple polysilicon technology”. In: *1984 International Electron Devices Meeting*. IEEE. 1984, pp. 464–467.

[38] Fujio Masuoka et al. “New ultra high density EPROM and flash EEPROM with NAND structure cell”. In: *1987 International Electron Devices Meeting*. IEEE. 1987, pp. 552–555.

[39] Ali Khakifirooz et al. “A 1.67 tb, 5b/cell flash memory fabricated in 192-layer floating gate 3d-nand technology and featuring a 23.3 gb/mm 2 bit density”. In: *IEEE Solid-State Circuits Letters* 6 (2023), pp. 161–164.

[40] Sayeef Salahuddin, Kai Ni, and Suman Datta. “The era of hyper-scaling in electronics”. In: *Nature electronics* 1.8 (2018), pp. 442–450.

[41] MF Gonzalez-Zalba et al. “Scaling silicon-based quantum computing using CMOS technology”. In: *Nature Electronics* 4.12 (2021), pp. 872–884.

[42] C Rinn Cleavelin, Billy C Covington, and Lawrence A Larson. “Front end of line considerations for progression beyond the 100 nm node ultrashallow junction requirements”. In: *Journal of Vacuum Science & Technology B: Microelectronics and Nanometer Structures Processing, Measurement, and Phenomena* 18.1 (2000), pp. 346–353.

[43] Jun Hwan Moon et al. “Materials quest for advanced interconnect metallization in integrated circuits”. In: *Advanced Science* 10.23 (2023), p. 2207321.

[44] Gordon E Moore et al. “Progress in digital integrated electronics”. In: *Electron devices meeting*. Vol. 21. Washington, DC. 1975, pp. 11–13.

[45] Suman Datta. “Recent advances in high performance CMOS transistors: From planar to non-planar”. In: *The Electrochemical Society Interface* 22.1 (2013), p. 41.

[46] Mark T Bohr and Ian A Young. “CMOS scaling trends and beyond”. In: *IEEE Micro* 37.6 (2017), pp. 20–29.

[47] Yuan Taur et al. "CMOS scaling into the nanometer regime". In: *Proceedings of the IEEE* 85.4 (1997), pp. 486–504.

[48] Qian Xie, Jun Xu, and Yuan Taur. "Review and critique of analytic models of MOSFET short-channel effects in subthreshold". In: *IEEE transactions on electron devices* 59.6 (2012), pp. 1569–1579.

[49] H Iwai et al. "Velocity saturation effect on short-channel MOS transistor capacitance". In: *IEEE electron device letters* 6.3 (1985), pp. 120–122.

[50] Ronald R Troutman. "VLSI limitations from drain-induced barrier lowering". In: *IEEE Journal of Solid-State Circuits* 14.2 (1979), pp. 383–391.

[51] Savvas G Chamberlain and Sannasi Ramanan. "Drain-induced barrier-lowering analysis in VSLI MOSFET devices using two-dimensional numerical simulations". In: *IEEE transactions on electron devices* 33.11 (1986), pp. 1745–1753.

[52] Tor A Fjeldly and Michael Shur. "Threshold voltage modeling and the subthreshold regime of operation of short-channel MOSFETs". In: *IEEE Transactions on Electron Devices* 40.1 (1993), pp. 137–145.

[53] Jian Chen et al. "Subbreakdown drain leakage current in MOSFET". In: *IEEE Electron Device Letters* 8.11 (1987), pp. 515–517.

[54] Xiaobin Yuan et al. "Gate-induced-drain-leakage current in 45-nm CMOS technology". In: *IEEE Transactions on Device and Materials Reliability* 8.3 (2008), pp. 501–508.

[55] Pranita Kerber et al. "GIDL in doped and undoped FinFET devices for low-leakage applications". In: *IEEE Electron Device Letters* 34.1 (2012), pp. 6–8.

[56] TY Chan et al. "The impact of gate-induced drain leakage current on MOSFET scaling". In: *1987 International Electron Devices Meeting*. IEEE. 1987, pp. 718–721.

[57] George K Celler and Sorin Cristoloveanu. "Frontiers of silicon-on-insulator". In: *Journal of Applied Physics* 93.9 (2003), pp. 4955–4978.

[58] Sorin Cristoloveanu. "Silicon on insulator technologies and devices: from present to future". In: *Solid-State Electronics* 45.8 (2001), pp. 1403–1411.

[59] Sorin Cristoloveanu and Sheng Li. *Electrical characterization of silicon-on-insulator materials and devices*. Vol. 305. Springer Science & Business Media, 2013.

[60] Martin M Frank et al. "Scaling the MOSFET gate dielectric: From high-k to higher-k?" In: *Microelectronic Engineering* 86.7-9 (2009), pp. 1603–1608.

[61] HR Huff et al. "High-k gate stacks for planar, scaled CMOS integrated circuits". In: *Microelectronic Engineering* 69.2-4 (2003), pp. 152–167.

[62] Gang He et al. "Integrations and challenges of novel high-k gate stacks in advanced CMOS technology". In: *Progress in Materials Science* 56.5 (2011), pp. 475–572.

[63] Micharl Riordan. "The silicon dioxide solution". In: *IEEE Spectrum* 44.12 (2007), pp. 51–56.

[64] Robert H Dennard et al. "Design of ion-implanted MOSFET's with very small physical dimensions". In: *IEEE Journal of solid-state circuits* 9.5 (1974), pp. 256–268.

[65] Mark Bohr. "A 30 year retrospective on Dennard's MOSFET scaling paper". In: *IEEE Solid-State Circuits Society Newsletter* 12.1 (2007), pp. 11–13.

[66] Toshihiro Sekigawa and Yasuhiro Hayashi. "Calculated threshold-voltage characteristics of an XMOS transistor having an additional bottom gate". In: *Solid-State Electronics* 27.8-9 (1984), pp. 827–828.

[67] Digh Hisamoto et al. "A fully depleted lean-channel transistor (DELTA)-a novel vertical ultra thin SOI MOSFET". In: *International Technical Digest on Electron Devices Meeting*. IEEE. 1989, pp. 833–836.

[68] Digh Hisamoto, Toru Kaga, and Elji Takeda. "Impact of the vertical SOI 'DELTA' structure on planar device technology". In: *IEEE Transactions on Electron Devices* 38.6 (1991), pp. 1419–1424.

[69] Digh Hisamoto et al. "A new stacked cell structure for giga-bit DRAMs using vertical ultra-thin SOI (DELTA) MOSFETs". In: *International Electron Devices Meeting 1991 [Technical Digest]*. IEEE. 1991, pp. 959–961.

[70] Digh Hisamoto et al. "A folded-channel MOSFET for deep-sub-tenth micron era". In: *IEDM Tech. Dig* 1998 (1998), pp. 1032–1034.

[71] Dick James. "Intel Ivy Bridge unveiled—The first commercial tri-gate, high-k, metal-gate CPU". In: *Proceedings of the IEEE 2012 Custom Integrated Circuits Conference*. IEEE. 2012, pp. 1–4.

[72] S Natarajan et al. "A 14nm logic technology featuring 2 nd-generation finfet, air-gapped interconnects, self-aligned double patterning and a 0.0588 μm 2 sram cell size". In: *2014 IEEE international electron devices meeting*. IEEE. 2014, pp. 3–7.

[73] Asharani Samal, Suman Lata Tripathi, and Sushanta Kumar Mohapatra. "A journey from bulk MOSFET to 3 nm and beyond". In: *Transactions on Electrical and Electronic Materials* 21.5 (2020), pp. 443–455.

[74] Laixiang Qin et al. "Recent developments in negative capacitance gate-all-around field effect transistors: a review". In: *IEEE Access* 11 (2023), pp. 14028–14042.

[75] H Takato et al. "High performance CMOS surrounding gate transistor (SGT) for ultra high density LSIs". In: *Technical Digest., International Electron Devices Meeting*. IEEE. 1988, pp. 222–225.

[76] Jean-Pierre Colinge et al. "Silicon-on-insulator'gate-all-around device'". In: *International technical digest on electron devices*. IEEE. 1990, pp. 595–598.

[77] Hyunjin Lee et al. "Sub-5nm all-around gate FinFET for ultimate scaling". In: *2006 Symposium on VLSI Technology, 2006. Digest of Technical Papers.* IEEE. 2006, pp. 58–59.

[78] N Singh et al. "High-performance fully depleted silicon nanowire (diameter/spl les/5 nm) gate-all-around CMOS devices". In: *IEEE Electron Device Letters* 27.5 (2006), pp. 383–386.

[79] N Loubet et al. "Stacked nanosheet gate-all-around transistor to enable scaling beyond FinFET". In: *2017 symposium on VLSI technology.* IEEE. 2017, T230–T231.

[80] Lun-Chun Chen et al. "High-performance stacked double-layer N-channel poly-Si nanosheet multigate thin-film transistors". In: *IEEE Electron Device Letters* 38.9 (2017), pp. 1256–1258.

[81] Yusuke Oniki, Efraín Altamirano-Sánchez, and Frank Holsteyns. "Selective etches for gate-all-around (GAA) device integration: Opportunities and challenges". In: *ECS Transactions* 92.2 (2019), p. 3.

[82] Sung-Young Lee et al. "A novel multibridge-channel MOSFET (MBCFET): fabrication technologies and characteristics". In: *IEEE transactions on nanotechnology* 2.4 (2003), pp. 253–257.

[83] Jaehun Jeong et al. "World's first GAA 3nm foundry platform technology (SF3) with novel multi-bridge-channel-FET (MBCFET™) process". In: *2023 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits).* IEEE. 2023, pp. 1–2.

[84] Doyoung Jang et al. "Device exploration of nanosheet transistors for sub-7-nm technology node". In: *IEEE Transactions on Electron Devices* 64.6 (2017), pp. 2707–2713.

[85] Anabela Veloso et al. "Nanowire & nanosheet FETs for ultra-scaled, high-density logic and memory applications". In: *Solid-State Electronics* 168 (2020), p. 107736.

[86] Sergey V Kalinin et al. "A New Revolution in Logic Silicon IC Technology: GAA FETs are Replacing FinFETs". In: *2023 IEEE XVI International Scientific and Technical Conference Actual Problems of Electronic Instrument Engineering (APEIE)*. IEEE. 2023, pp. 160–165.

[87] Owain Vaughan. "Two nanometre CMOS technology". In: *Nature Electronics* (2024), pp. 1–1.

[88] Shubo Zhang. "Review of modern field effect transistor technologies for scaling". In: *Journal of Physics: Conference Series*. Vol. 1617. 1. IOP Publishing. 2020, p. 012054.

[89] Wei Hu and Feng Li. "Scaling beyond 7nm node: An overview of gate-all-around fets". In: *2021 9th international symposium on next generation electronics (ISNE)*. IEEE. 2021, pp. 1–6.

[90] S Subramanian et al. "First monolithic integration of 3d complementary fet (cfet) on 300mm wafers". In: *2020 Ieee Symposium on Vlsi Technology*. IEEE. 2020, pp. 1–2.

[91] Xusheng Wu et al. "A three-dimensional stacked Fin-CMOS technology for high-density ULSI circuits". In: *IEEE transactions on electron devices* 52.9 (2005), pp. 1998–2003.

[92] Kelin J Kuhn et al. *Silicon and silicon germanium nanowire structures*. US Patent 9,129,829. Sept. 2015.

[93] Julien Ryckaert et al. "The Complementary FET (CFET) for CMOS scaling beyond N3". In: *2018 IEEE Symposium on Vlsi Technology*. IEEE. 2018, pp. 141–142.

[94] Shixin Li et al. "Vertically Stacked Nanosheet Number Optimization Strategy for Complementary FET (CFET) Scaling Beyond 2 nm". In: *IEEE Transactions on Electron Devices* (2023).

[95] B Vincent et al. "A benchmark study of complementary-field effect transistor (CFET) process integration options done by virtual fabrication". In: *IEEE Journal of the Electron Devices Society* 8 (2020), pp. 668–673.

[96] Eunbin Park and Taigon Song. "Complementary FET (CFET) standard cell design for low parasitics and its impact on VLSI prediction at 3-nm process". In: *IEEE Transactions on Very Large Scale Integration (VLSI) Systems* 31.2 (2022), pp. 177–187.

[97] Wei-Cheng Kang et al. "A complementary FET (CFET)-based NAND design to reduce RC delay". In: *IEEE Electron Device Letters* 43.5 (2022), pp. 678–681.

[98] Krithika Dhananjay et al. "Monolithic 3D Integrated circuits: Recent trends and future prospects". In: *IEEE Transactions on Circuits and Systems II: Express Briefs* 68.3 (2021), pp. 837–843.

[99] John U Knickerbocker et al. "3D silicon integration". In: *2008 58th Electronic Components and Technology Conference*. IEEE. 2008, pp. 538–543.

[100] Joonyoung Kim and Younsu Kim. "HBM: Memory solution for bandwidth-hungry processors". In: *2014 IEEE Hot Chips 26 Symposium (HCS)*. IEEE. 2014, pp. 1–24.

[101] Hongshin Jun et al. "Hbm (high bandwidth memory) dram technology and architecture". In: *2017 IEEE International Memory Workshop (IMW)*. IEEE. 2017, pp. 1–4.

[102] Seung Wook Yoon et al. "3D TSV processes and its assembly/packaging technology". In: *2009 IEEE International Conference on 3D System Integration*. IEEE. 2009, pp. 1–5.

[103] John H Lau. "Overview and outlook of through-silicon via (TSV) and 3D integrations". In: *Microelectronics International* 28.2 (2011), pp. 8–22.

[104] Meng-Jen Wang et al. "TSV technology for 2.5 D IC solution". In: *2012 IEEE 62nd Electronic Components and Technology Conference*. IEEE. 2012, pp. 284–288.

[105] Kwiwook Kim and Myeong-jae Park. "Present and Future, Challenges of High Bandwith Memory (HBM)". In: *2024 IEEE International Memory Workshop (IMW)*. IEEE. 2024, pp. 1–4.

[106] Li Li et al. "3D SiP with organic interposer for ASIC and memory integration". In: *2016 IEEE 66th Electronic Components and Technology Conference (ECTC)*. IEEE. 2016, pp. 1445–1450.

[107] Mindy D Bishop et al. "Monolithic 3-D integration". In: *IEEE Micro* 39.6 (2019), pp. 16–27.

[108] T Naito et al. "World's first monolithic 3D-FPGA with TFT SRAM over 90nm 9 layer Cu CMOS". In: *2010 Symposium on VLSI Technology*. IEEE. 2010, pp. 219–220.

[109] A Belmonte et al. "Capacitor-less, long-retention (¿ 400s) DRAM cell paving the way towards low-power and high-density monolithic 3D DRAM". In: *2020 IEEE International Electron Devices Meeting (IEDM)*. IEEE. 2020, pp. 28–2.

[110] Takeya Hirose et al. "1-Mbit 3D DRAM Using a Monolithically Stacked Structure of a Si CMOS and Heterogeneous IGZO FETs". In: *IEEE Journal of the Electron Devices Society* (2024).

[111] Yuanqing Cheng, Xiaochen Guo, and Vasilis F Pavlidis. "Emerging monolithic 3D integration: Opportunities and challenges from the computer system perspective". In: *Integration* 85 (2022), pp. 97–107.

[112] Stanford R Ovshinsky. "Reversible electrical switching phenomena in disordered structures". In: *Physical review letters* 21.20 (1968), p. 1450.

[113] RG Neale and John A Aseltine. "The application of amorphous materials to computer memories". In: *IEEE Transactions on Electron Devices* 20.2 (1973), pp. 195–205.

[114] Stefan Lai. "Current status of the phase change memory and its future". In: *IEEE International Electron Devices Meeting 2003.* IEEE. 2003, pp. 10–1.

[115] MJ Kang et al. "PRAM cell technology and characterization in 20nm node size". In: *2011 International Electron Devices Meeting.* IEEE. 2011, pp. 3–1.

[116] Geoffrey W Burr et al. "Phase change memory technology". In: *Journal of Vacuum Science & Technology B* 28.2 (2010), pp. 223–262.

[117] Geoffrey W Burr et al. "Neuromorphic computing using non-volatile memory". In: *Advances in Physics: X* 2.1 (2017), pp. 89–124.

[118] Evangelos Eleftheriou et al. "Deep learning acceleration based on in-memory computing". In: *IBM Journal of Research and Development* 63.6 (2019), pp. 7–1.

[119] Manuel Le Gallo and Abu Sebastian. "An overview of phase-change memory device physics". In: *Journal of Physics D: Applied Physics* 53.21 (2020), p. 213002.

[120] Manuel Le Gallo et al. "A 64-core mixed-signal in-memory compute chip based on phase-change memory for deep neural network inference". In: *Nature Electronics* 6.9 (2023), pp. 680–693.

[121] H-S Philip Wong et al. "Phase change memory". In: *Proceedings of the IEEE* 98.12 (2010), pp. 2201–2227.

[122] Geoffrey W Burr et al. "Recent progress in phase-change memory technology". In: *IEEE Journal on Emerging and Selected Topics in Circuits and Systems* 6.2 (2016), pp. 146–162.

[123] Scott W Fong, Christopher M Neumann, and H-S Philip Wong. "Phase-change memory—Towards a storage-class memory". In: *IEEE Transactions on Electron Devices* 64.11 (2017), pp. 4374–4385.

[124] Taehoon Kim and Seungyun Lee. "Evolution of phase-change memory for the storage-class memory and beyond". In: *IEEE Transactions on Electron Devices* 67.4 (2020), pp. 1394–1406.

[125] *Numonyx Introduces New Phase Change Memory Devices.* `https://investors.micron.com/news-releases/news-release-details/numonyx-introduces-new-phase-change-memory-devices`. [Accessed 07-01-2025].

[126] *Intel® Optane™ Persistent Memory (PMem) — intel.com.* `https://www.intel.com/content/www/us/en/products/details/memory-storage/optane-dc-persistent-memory.html`. [Accessed 07-01-2025].

[127] F Xiong et al. "Towards ultimate scaling limits of phase-change memory". In: *2016 IEEE International Electron Devices Meeting (IEDM)*. IEEE. 2016, pp. 4–1.

[128] Paolo Fantini. "Phase change memory applications: the history, the present and the future". In: *Journal of Physics D: Applied Physics* 53.28 (2020), p. 283002.

[129] Joseph Valasek. "Piezo-electric and allied phenomena in Rochelle salt". In: *Physical review* 17.4 (1921), p. 475.

[130] JL Moll and YJIToED Tarui. "A new solid state memory resistor". In: *IEEE Transactions on Electron Devices* 10.5 (1963), pp. 338–338.

[131] Shu-Yau Wu. "A new ferroelectric memory device, metal-ferroelectric-semiconductor transistor". In: *IEEE Transactions on Electron Devices* 21.8 (1974), pp. 499–504.

[132] Lane W Martin and Andrew M Rappe. "Thin-film ferroelectric materials and their applications". In: *Nature Reviews Materials* 2.2 (2016), pp. 1–14.

[133] T Francois et al. "Demonstration of BEOL-compatible ferroelectric Hf 0.5 Zr 0.5 O 2 scaled FeRAM co-integrated with 130nm CMOS for embedded NVM applications". In: *2019 IEEE International Electron Devices Meeting (IEDM)*. IEEE. 2019, pp. 15–7.

[134] Jae Young Kim, Min-Ju Choi, and Ho Won Jang. "Ferroelectric field effect transistors: Progress and perspective". In: *APL Materials* 9.2 (2021).

[135] Wonho Lee et al. "Flexible graphene–PZT ferroelectric nonvolatile memory". In: *Nanotechnology* 24.47 (2013), p. 475202.

[136] Mohamed T Ghoneim et al. "Thin PZT-based ferroelectric capacitors on flexible silicon for nonvolatile memory applications". In: *Advanced electronic materials* 1.6 (2015), p. 1500045.

[137] NK Kim et al. "(Bi, La) 4Ti3O12 (BLT) thin films grown from nanocrystalline perovskite nuclei for ferroelectric memory devices". In: *Applied physics letters* 85.18 (2004), pp. 4118–4120.

[138] Qiao Jin et al. "Enhanced resistive memory in Nb-doped BaTiO3 ferroelectric diodes". In: *Applied Physics Letters* 111.3 (2017).

[139] Orlando Auciello. "A critical comparative review of PZT and SBT-based science and technology for non-volatile ferroelectric memories". In: *Integrated Ferroelectrics* 15.1-4 (1997), pp. 211–220.

[140] Ludovic Goux et al. "A highly reliable 3-D integrated SBT ferroelectric capacitor enabling FeRAM scaling". In: *IEEE transactions on electron devices* 52.4 (2005), pp. 447–453.

[141] Seok Ju Kang et al. "Printable ferroelectric PVDF/PMMA blend films with ultralow roughness for low voltage non-volatile polymer memory". In: *Advanced Functional Materials* 19.17 (2009), pp. 2812–2818.

[142] Youn Jung Park et al. "Ordered ferroelectric PVDF- TrFE thin films by high throughput epitaxy for nonvolatile polymer memory". In: *Macromolecules* 41.22 (2008), pp. 8648–8654.

[143] Daisaburo Takashima. "Overview of FeRAMs: Trends and perspectives". In: *2011 11th Annual Non-Volatile Memory Technology Symposium Proceeding*. IEEE. 2011, pp. 1–6.

[144] Matthew Dawber, KM Rabe, and JF Scott. "Physics of thin-film ferroelectric oxides". In: *Reviews of modern physics* 77.4 (2005), pp. 1083–1130.

[145] T Mikolajick, U Schroeder, and S Slesazeck. "The past, the present, and the future of ferroelectric memories". In: *IEEE Transactions on Electron Devices* 67.4 (2020), pp. 1434–1443.

[146] Infineon Technologies AG. *EXCELON™ F-RAM - Infineon Technologies — infineon.com*. `https://www.infineon.com/cms/en/product/memories/f-ram-ferroelectric-ram/excelon-f-ram/`. [Accessed 07-01-2025].

[147] *FeRAM Product List RAMXEED — ramxeed.com*. `https://www.ramxeed.com/products/feram/feram-products.html`. [Accessed 02-01-2025].

[148] JF Gibbons and WE Beadle. "Switching properties of thin NiO films". In: *Solid-State Electronics* 7.11 (1964), pp. 785–790.

[149] PH Nielsen and NM Bashara. "The reversible voltage-induced initial resistance in the negative resistance sandwich structure". In: *IEEE Transactions on Electron Devices* 11.5 (1964), pp. 243–244.

[150] Masayoshi Nakayama. "ReRAM technologies: applications and outlook". In: *2017 IEEE International Memory Workshop (IMW)*. IEEE. 2017, pp. 1–4.

[151] Yangyin Chen. "ReRAM: History, status, and future". In: *IEEE Transactions on Electron Devices* 67.4 (2020), pp. 1420–1433.

[152] WW Zhuang et al. "Novel colossal magnetoresistive thin film nonvolatile resistance random access memory (RRAM)". In: *Digest. International Electron Devices Meeting,* IEEE. 2002, pp. 193–196.

[153] Bogdan Govoreanu et al. "10× 10nm 2 Hf/HfO x crossbar resistive RAM with excellent performance, reliability and low-energy operation". In: *2011 International Electron Devices Meeting*. IEEE. 2011, pp. 31–6.

[154] Yukio Hayakawa et al. "Highly reliable TaO x ReRAM with centralized filament for 28-nm embedded application". In: *2015 Symposium on VLSI Technology (VLSI Technology)*. IEEE. 2015, T14–T15.

[155] Jury Sandrini et al. "Co-design of ReRAM passive crossbar arrays integrated in 180 nm CMOS technology". In: *IEEE Journal on Emerging and Selected Topics in Circuits and Systems* 6.3 (2016), pp. 339–351.

[156] *ReRAM Product List RAMXEED — ramxeed.com*. `https://www.ramxeed.com/products/reram/reram-products.html`. [Accessed 02-01-2025].

[157] Sabpreet Bhatti et al. "Spintronics based random access memory: a review". In: *Materials Today* 20.9 (2017), pp. 530–548.

[158] Bernard Dieny et al. "Opportunities and challenges for spintronics in the microelectronics industry". In: *Nature Electronics* 3.8 (2020), pp. 446–459.

[159] Mario Norberto Baibich et al. "Giant magnetoresistance of (001) Fe/(001) Cr magnetic superlattices". In: *Physical review letters* 61.21 (1988), p. 2472.

[160] Grünberg Binasch et al. "Enhanced magnetoresistance in layered magnetic structures with antiferromagnetic interlayer exchange". In: *Physical review B* 39.7 (1989), p. 4828.

[161] Terunobu Miyazaki and Nobuki Tezuka. "Giant magnetic tunneling effect in Fe/Al2O3/Fe junction". In: *Journal of magnetism and magnetic materials* 139.3 (1995), pp. L231–L234.

[162] Jagadeesh Subbaiah Moodera et al. "Large magnetoresistance at room temperature in ferromagnetic thin film tunnel junctions". In: *Physical review letters* 74.16 (1995), p. 3273.

[163] John C Slonczewski. "Current-driven excitation of magnetic multilayers". In: *Journal of Magnetism and Magnetic Materials* 159.1-2 (1996), pp. L1–L7.

[164] Luc Berger. "Emission of spin waves by a magnetic multilayer traversed by a current". In: *Physical Review B* 54.13 (1996), p. 9353.

[165] Suock Chung et al. "Fully integrated 54nm STT-RAM with the smallest bit cell dimension for high density memory application". In: *2010 International Electron Devices Meeting*. IEEE. 2010, pp. 12–7.

[166] Yong Kyu Lee et al. "Embedded STT-MRAM in 28-nm FDSOI logic process for industrial MCU/IoT application". In: *2018 IEEE Symposium on VLSI Technology*. IEEE. 2018, pp. 181–182.

[167] Ioan Mihai Miron et al. "Perpendicular switching of a single ferromagnetic layer induced by in-plane current injection". In: *Nature* 476.7359 (2011), pp. 189–193.

[168] Kevin Garello et al. "SOT-MRAM 300mm integration for low power and ultrafast embedded memories". In: *2018 IEEE symposium on VLSI Circuits*. IEEE. 2018, pp. 81–82.

[169] Kevin Garello et al. "Manufacturable 300mm platform solution for field-free switching SOT-MRAM". In: *2019 Symposium on VLSI Circuits*. IEEE. 2019, T194–T195.

[170] Hyunsoo Yang et al. "Two-dimensional materials prospects for non-volatile spintronic memories". In: *Nature* 606.7915 (2022), pp. 663–673.

[171] *Spin-transfer Torque DDR Products — Everspin — everspin.com*. `https://www.everspin.com/spin-transfer-torque-ddr-products`. [Accessed 07-01-2025].

[172] *MRAM for Aerospace Applications: Reliable and Radiation-Hardened — avalanche-technology.com*. `https://www.avalanche-technology.com/products/discrete-mram/boot/`. [Accessed 07-01-2025].

[173] Oleg Golonzka et al. "MRAM as embedded non-volatile memory solution for 22FFL FinFET technology". In: *2018 IEEE International Electron Devices Meeting (IEDM)*. IEEE. 2018, pp. 18–1.

[174] YJ Song et al. "Demonstration of highly manufacturable STT-MRAM embedded in 28nm logic". In: *2018 IEEE International Electron Devices Meeting (IEDM)*. IEEE. 2018, pp. 18–2.

[175] *Samsung Electronics Starts Commercial Shipment of eMRAM Product Based on 28nm FD-SOI Process — Samsung Semiconductor Global — semiconductor.samsung.com*. `https://semiconductor.samsung.com/news-events/news/samsung-electronics-starts-commercial-shipment-of-emram-product-based-on-28nm-fd-soi-process/`. [Accessed 02-01-2025].

[176] Eric Millington. *Making New Memories: 22nm eMRAM is Ready to Displace eFlash — GlobalFoundries — gf.com*. `https://gf.com/blog/making-new-memories-22nm-emram-ready-displace-eflash/`. [Accessed 02-01-2025].

[177] Dmytro Apalkov, Bernard Dieny, and Jon M Slaughter. "Magnetoresistive random access memory". In: *Proceedings of the IEEE* 104.10 (2016), pp. 1796–1830.

[178] Tetsuo Endoh et al. "Recent progresses in STT-MRAM and SOT-MRAM for next generation MRAM". In: *2020 IEEE Symposium on VLSI Technology*. IEEE. 2020, pp. 1–2.

[179] I-M Park et al. "Enhanced Endurance Characteristics in High Performance 16nm Selector Only Memory (SOM)". In: *2023 International Electron Devices Meeting (IEDM)*. IEEE. 2023, pp. 1–4.

[180] Abu Sebastian et al. "Memory devices and applications for in-memory computing". In: *Nature nanotechnology* 15.7 (2020), pp. 529–544.

[181] John Backus. "Can programming be liberated from the von Neumann style? A functional style and its algebra of programs". In: *Communications of the ACM* 21.8 (1978), pp. 613–641.

[182] Wm A Wulf and Sally A McKee. "Hitting the memory wall: Implications of the obvious". In: *ACM SIGARCH computer architecture news* 23.1 (1995), pp. 20–24.

[183] Onur Mutlu et al. "Processing data where it makes sense: Enabling in-memory computation". In: *Microprocessors and Microsystems* 67 (2019), pp. 28–41.

[184] Vishal Sharma, Hyunjoon Kim, and Tony Tae-Hyoung Kim. "A 64 Kb reconfigurable full-precision digital ReRAM-based compute-in-memory for artificial intelligence applications". In: *IEEE Transactions on Circuits and Systems I: Regular Papers* 69.8 (2022), pp. 3284–3296.

[185] Sai Zhang, Kejie Huang, and Haibin Shen. "A robust 8-bit non-volatile computing-in-memory core for low-power parallel MAC operations". In: *IEEE Transactions on Circuits and Systems I: Regular Papers* 67.6 (2020), pp. 1867–1880.

[186] Qi Liu et al. "33.2 A fully integrated analog ReRAM based 78.4 TOPS/W compute-in-memory chip with fully parallel MAC computing". In: *2020 IEEE International Solid-State Circuits Conference-(ISSCC)*. IEEE. 2020, pp. 500–502.

[187] Naveen Verma et al. "In-memory computing: Advances and prospects". In: *IEEE solid-state circuits magazine* 11.3 (2019), pp. 43–55.

[188] Jintao Zhang, Zhuo Wang, and Naveen Verma. "In-memory computation of a machine-learning classifier in a standard 6T SRAM array". In: *IEEE Journal of Solid-State Circuits* 52.4 (2017), pp. 915–924.

[189] Chuan-Jia Jhang et al. "Challenges and trends of SRAM-based computing-in-memory for AI edge devices". In: *IEEE Transactions on Circuits and Systems I: Regular Papers* 68.5 (2021), pp. 1773–1786.

[190] Sangjin Kim and Hoi-Jun Yoo. "An Overview of Computing-in-Memory Circuits with DRAM and NVM". In: *IEEE Transactions on Circuits and Systems II: Express Briefs* (2023).

[191] Daniele Ielmini and H-S Philip Wong. "In-memory computing with resistive switching devices". In: *Nature electronics* 1.6 (2018), pp. 333–343.

[192] Yu Pan et al. "A multilevel cell STT-MRAM-based computing in-memory accelerator for binary convolutional neural network". In: *IEEE Transactions on Magnetics* 54.11 (2018), pp. 1–5.

[193] Hao Cai et al. "Proposal of analog in-memory computing with magnified tunnel magnetoresistance ratio and universal STT-MRAM cell". In: *IEEE Transactions on Circuits and Systems I: Regular Papers* 69.4 (2022), pp. 1519–1531.

[194] Daniele Ielmini and Giacomo Pedretti. "Device and circuit architectures for in-memory computing". In: *Advanced Intelligent Systems* 2.7 (2020), p. 2000040.

[195] R Stanley Williams. "What's Next?[The end of Moore's law]". In: *Computing in Science & Engineering* 19.2 (2017), pp. 7–13.

[196] Peng Yao et al. "Fully hardware-implemented memristor convolutional neural network". In: *Nature* 577.7792 (2020), pp. 641–646.

[197] Shimeng Yu and Pai-Yu Chen. "Emerging memory technologies: Recent trends and prospects". In: *IEEE Solid-State Circuits Magazine* 8.2 (2016), pp. 43–56.

[198] Kirk Prall. "Benchmarking and metrics for emerging memory". In: *2017 IEEE International Memory Workshop (IMW)*. IEEE. 2017, pp. 1–5.

[199] Yiran Chen et al. "Recent technology advances of emerging memories". In: *IEEE Design & Test* 34.3 (2017), pp. 8–22.

[200] Meenatchi Jagasivamani et al. "Design for ReRAM-based main-memory architectures". In: *Proceedings of the International Symposium on Memory Systems*. 2019, pp. 342–350.

[201] Writam Banerjee. "Challenges and applications of emerging nonvolatile memory devices". In: *Electronics* 9.6 (2020), p. 1029.

[202] Yole Intelligence. *Status of the Memory Industry 2024*. `https://www.yolegroup.com/product/report/status-of-the-memory-industry-2024/`. [Accessed 08-01-2025].

[203] ER Brown. "Resonant tunneling in high-speed double barrier". In: *Hot Carriers in Semiconductor Nanostructures: Physics and Applications* 469 (2012).

[204] L_L Chang, Leo Esaki, and R Tsu. "Resonant tunneling in semiconductor double barriers". In: *Applied physics letters* 24.12 (1974), pp. 593–595.

[205] R Tsu and Leo Esaki. "Tunneling in a finite superlattice". In: *Applied Physics Letters* 22.11 (1973), pp. 562–564.

[206] LD Macks et al. "Resonant tunneling in double-quantum-well triple-barrier heterostructures". In: *Physical Review B* 54.7 (1996), p. 4857.

[207] Philip Derek Buckle et al. "Charge accumulation in GaAs/AlGaAs triple barrier resonant tunneling structures". In: *Journal of applied physics* 83.2 (1998), pp. 882–887.

[208] PM Martin et al. "Magnetic-field-induced resonance in a triple-barrier resonant tunnelling diode". In: *Semiconductor Science and Technology* 9.5S (1994), p. 493.

[209] Taotao Rong et al. "Theoretical modeling of triple-barrier resonant-tunneling diodes based on AlGaN/GaN heterostructures". In: *physica status solidi (a)* 216.23 (2019), p. 1900471.

[210] Kinichiro Araki. "Analysis of barrier transmission in resonant tunneling diodes". In: *Journal of applied physics* 62.3 (1987), pp. 1059–1069.

[211] Timothy B Boykin. "Approximations for the resonant-tunneling diode current: Implications for triple-barrier devices". In: *Journal of applied physics* 78.11 (1995), pp. 6818–6821.

[212] Gyungock Kim, Kwang Man Koh, and Chong Hoon Kim. "Evidence of the enhanced resonant tunneling effect in a triple-barrier heterostructure". In: *Superlattices and Microstructures* 29.1 (2001), pp. 51–55.

[213] JM Xu, VV Malov, and LV Iogansen. "Comparison of resonant tunneling in a double-quantum-well three-barrier system and a single-quantum-well double-barrier system". In: *Physical Review B* 47.12 (1993), p. 7253.

[214] Dominic Lane and Manus Hayne. "Simulations of resonant tunnelling through InAs/AlSb heterostructures for ULTRARAM™ memory". In: *Journal of Physics D: Applied Physics* 54.35 (2021), p. 355104.

[215] Clint B Geller et al. "Computational band-structure engineering of III–V semiconductor alloys". In: *Applied Physics Letters* 79.3 (2001), pp. 368–370.

[216] Wl/odzimierz Nakwaski. "Thermal conductivity of binary, ternary, and quaternary III-V compounds". In: *Journal of Applied Physics* 64.1 (1988), pp. 159–166.

[217] Thomas F Kuech. "III-V compound semiconductors: Growth and structures". In: *Progress in crystal growth and characterization of materials* 62.2 (2016), pp. 352–370.

[218] D Chattopadhyay, SK Sutradhar, and BR Nag. "Electron transport in direct-gap III-V ternary alloys". In: *Journal of Physics C: Solid State Physics* 14.6 (1981), p. 891.

[219] Abdelkader Aissat et al. "Modeling of Ga1- xInxAs1- y- zNySbz/GaAs quantum well properties for near-infrared lasers". In: *Materials science in semiconductor processing* 16.6 (2013), pp. 1936–1942.

[220] Jesús A Del Alamo. "Nanometre-scale electronics with III–V compound semiconductors". In: *Nature* 479.7373 (2011), pp. 317–323.

[221] Raseong Kim, Uygar E Avci, and Ian A Young. "Comprehensive performance benchmarking of III-V and Si nMOSFETs (gate length= 13 nm) considering supply voltage and OFF-current". In: *IEEE Transactions on Electron Devices* 62.3 (2015), pp. 713–721.

[222] Jesús A Del Alamo et al. "Nanometer-Scale III-V MOSFETs". In: *IEEE Journal of the Electron Devices Society* 4.5 (2016), pp. 205–214.

[223] GA Antypas and TO Yep. "Growth and Characterization of Liquid-Phase Epitaxial InAs1- x P x". In: *Journal of Applied Physics* 42.8 (1971), pp. 3201–3204.

[224] I Vurgaftman, JR Meyer, and LR Ram-Mohan. "Applied Physics Review". In: *J. Appl. Phys* 89 (2001), p. 5815.

[225] J Devenson et al. "InAs/ AlSb quantum cascade lasers emitting below 3μm". In: *Applied physics letters* 90.11 (2007).

[226] Timothy B Boykin. "Current-voltage calculations for InAs/AlSb resonant-tunneling diodes". In: *Physical Review B* 51.7 (1995), p. 4289.

[227] Herbert Kroemer. "The 6.1 A family (InAs, GaSb, AlSb) and its heterostructures: a selective review". In: *Physica E: Low-dimensional Systems and Nanostructures* 20.3-4 (2004), pp. 196–203.

[228] Kanguk Kim et al. "14nm DRAM development and manufacturing". In: *2023 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)*. IEEE. 2023, pp. 1–2.

[229] Akira Goda. "Recent progress on 3D NAND flash technologies". In: *Electronics* 10.24 (2021), p. 3156.

[230] Andrea C Levi and Miroslav Kotrla. "Theory and simulation of crystal growth". In: *Journal of Physics: Condensed Matter* 9.2 (1997), p. 299.

[231] Tom Wilson. "The Design, Optimisation, and Characterisation of GaSb/GaAs Quantum Ring-Based Vertical-Cavity Devices Emitting at Telecoms Wavelengths." English. PhD thesis. Lancaster University, Feb. 2022. DOI: 10.17635/lancaster/thesis/1550.

[232] L Bolla. *GitHub - lbolla/EMpy: Electromagnetic Python — github.com*. https://github.com/lbolla/EMpy. [Accessed 18-12-2024]. 2017.

[233] Bo Gong and Gregory N Parsons. "Quantitative in situ infrared analysis of reactions between trimethylaluminum and polymers during Al 2 O 3 atomic layer deposition". In: *Journal of Materials Chemistry* 22.31 (2012), pp. 15672–15682.

[234] Zeyang Ren et al. "High temperature (300 C) ALD grown Al2O3 on hydrogen terminated diamond: Band offset and electrical properties of the MOSFETs". In: *Applied Physics Letters* 116.1 (2020).

[235] Donald L Smith, Andrew S Alimonda, and Frederick J von Preissig. "Mechanism of SiN x H y deposition from N2–SiH4 plasma". In: *Journal of Vacuum Science & Technology B: Microelectronics Processing and Phenomena* 8.3 (1990), pp. 551–557.

[236] Hideki Hasegawa. "Fermi level pinning and Schottky barrier height control at metal-semiconductor interfaces of InP and related materials". In: *Japanese Journal of Applied Physics* 38.2S (1999), p. 1098.

[237] *nextnano - Software for semiconductor nanodevices — nextnano.com*. https://www.nextnano.com/. [Accessed 09-01-2025].

[238] JW Lee et al. "Inductively coupled plasma etching of III–V semiconductors in Cl2-based chemistries". In: *Materials Science in Semiconductor Processing* 1.1 (1998), pp. 65–73.

[239] A Piotrowska and E Kaminska. "Ohmic contacts to III–V compound semiconductors". In: *Thin solid films* 193 (1990), pp. 511–527.

[240] LÖ Olsson et al. "Charge accumulation at InAs surfaces". In: *Physical review letters* 76.19 (1996), p. 3626.

[241] S Bhargava et al. "Fermi-level pinning position at the Au–InAs interface determined using ballistic electron emission microscopy". In: *Applied physics letters* 70.6 (1997), pp. 759–761.

[242] Suyeon Kim et al. "Influence of growth temperature on dielectric strength of Al2O3 thin films prepared via atomic layer deposition at low temperature". In: *Scientific Reports* 12.1 (2022), p. 5124.

[243] Robert H Doremus. "Diffusion in alumina". In: *Journal of applied physics* 100.10 (2006).

[244] Brian R Bennett et al. "Modulation doping of InAs/AlSb quantum wells using remote InAs donor layers". In: *Applied physics letters* 72.10 (1998), pp. 1193–1195.

[245] Yanbo Li, Yang Zhang, and Yiping Zeng. "Electron mobility in modulation-doped AlSb/InAs quantum wells". In: *Journal of Applied Physics* 109.7 (2011).

[246] JH Klootwijk and CE Timmering. "Merits and limitations of circular TLM structures for contact resistance determination for novel III-V HBTs". In: *Proceedings of the 2004 International Conference on Microelectronic Test Structures (IEEE Cat. No. 04CH37516)*. IEEE. 2004, pp. 247–252.

[247] Tae-Woo Kim, Dae-Hyun Kim, and Jesus A Del Alamo. "Logic characteristics of 40 nm thin-channel InAs HEMTs". In: *2010 22nd International Conference on Indium Phosphide and Related Materials (IPRM)*. IEEE. 2010, pp. 1–4.

[248] Aurelia Trevisan et al. "Defect formation in InGaAs/AlSb/InAs memory devices". In: *Journal of Vacuum Science & Technology B* 41.4 (2023).

[249] JR Söderström et al. "Growth and characterization of high current density, high-speed InAs/AlSb resonant tunneling diodes". In: *Applied physics letters* 58.3 (1991), pp. 275–277.

[250] John J Pekarik, Herbert Kroemer, and John H English. "An AlSb–InAs–AlSb double-heterojunction P-n-P bipolar transistor". In: *Journal of Vacuum Science & Technology B: Microelectronics and Nanometer Structures Processing, Measurement, and Phenomena* 10.2 (1992), pp. 1032–1034.

[251] Brian R Bennett et al. "Antimonide-based compound semiconductors for electronic devices: A review". In: *Solid-State Electronics* 49.12 (2005), pp. 1875–1895.

[252] R Magno et al. "The effect of defects on InAs/AlSb/GaSb resonant interband tunneling diodes". In: *Conference Proceedings. 2000 International Conference on Indium Phosphide and Related Materials (Cat. No. 00CH37107)*. IEEE. 2000, pp. 122–125.

[253] Gary Tuttle, Herbert Kroemer, and John H English. "Effects of interface layer sequencing on the transport properties of InAs/AlSb quantum wells: Evidence for antisite donors at the InAs/AlSb interface". In: *Journal of Applied Physics* 67.6 (1990), pp. 3032–3037.

[254] C Thomas et al. "High-mobility InAs 2DEGs on GaSb substrates: A platform for mesoscopic quantum transport". In: *Physical Review Materials* 2.10 (2018), p. 104602.

[255] Zhi Hua Li et al. "Buffer influence on AlSb/InAs/AlSb quantum wells". In: *Journal of crystal growth* 301 (2007), pp. 181–184.

[256] X Du et al. "Defect-related surface currents in InAs-based nBn infrared detectors". In: *Journal of Applied Physics* 123.21 (2018).

[257] F Frost et al. "Ion beam etching induced structural and electronic modification of InAs and InSb surfaces studied by Raman spectroscopy". In: *Journal of applied physics* 85.12 (1999), pp. 8378–8385.

[258] Berinder Brar and Herbert Kroemer. "Hole transport across the (Al, Ga)(As, Sb) barrier in InAs–(Al, Ga)(As, Sb) heterostructures". In: *Journal of applied physics* 83.2 (1998), pp. 894–899.

[259] Corentin Grillet, Alessandro Cresti, and Marco G Pala. "Vertical GaSb/AlSb/InAs heterojunction tunnel-FETs: a full quantum study". In: *IEEE Transactions on Electron Devices* 65.7 (2018), pp. 3038–3044.

[260] KF Longenbach, LF Luo, and WI Wang. "Resonant interband tunneling in InAs/GaSb/AlSb/InAs and GaSb/InAs/AlSb/GaSb heterostructures". In: *Applied physics letters* 57.15 (1990), pp. 1554–1556.

[261] JL Padilla et al. "Confinement-induced InAs/GaSb heterojunction electron–hole bilayer tunneling field-effect transistor". In: *Applied Physics Letters* 112.18 (2018).

[262] Mikhail N Polyanskiy. "Refractiveindex. info database of optical constants". In: *Scientific Data* 11.1 (2024), p. 94.