Forensic Applications of Atomic Force Microscopy

Daniel I. Konopinski

A dissertation submitted in partial fulfilment

of the requirements for the degree of

Doctor of Philosophy

of

University College London

Department of Electronic and Electrical Engineering,

University College London

August 2013

I, Daniel Icek Konopinski, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

The first project undertaken was to develop a currently non-existent forensic technique – data recovery from damaged SIM cards. SIM cards hold data valuable to a forensic investigator within non-volatile EEPROM/flash memory arrays. This data has been proven to be able to withstand temperatures up to 500°C, surviving such scenarios as house fires or criminal evidence disposal. A successful forensically-sound sample extraction, mounting and backside processing methodology was developed to expose the underside of a microcontroller circuit's floating gate transistor tunnel oxide, allowing probing via AFM-based electrical scanning probe techniques. Scanning Kelvin probe microscopy has thus far proved capable of detecting the presence of stored charge within the floating gates beneath the thin tunnel oxide layer, to the point of generating statistical distributions reflecting the threshold voltage states of the transistors.

The second project covered the novel forensic application of AFM as a complimentary technique to SEM examination of quartz grain surface textures. The analysis and interpretation of soil/sediment samples can provide indications of their provenance, and enable exclusionary comparisons to be made between samples pertinent to a forensic investigation. Multiple grains from four distinct sample sets were examined with the AFM, and various statistical figures of merit were derived. Canonical discriminant analysis was used to assess the discriminatory abilities of these statistical variables to better characterise the use of AFM results for grain classification. The final functions correctly classified 65.3% of original grouped cases, with the first 3 discriminant functions used in the analysis (Wilks' Lambda=0.336, p=0.000 < 0.01). This degree of discrimination shows a great deal of promise for the AFM as a quantitative corroborative technique to traditional SEM grain surface examination.

Acknowledgements

My deepest gratitude is accorded to the following people. My supervisor, Tony Kenyon, for all his support and guidance throughout this PhD; and more recently, for always taking the time to proof read my work (which I undoubtedly foisted on him at probably inopportune times). Steve Hudziak, for his patience, helpfulness, and above all else, friendship. I think I speak for all the regulars of the 9th floor 'nano' lab when I say that none of us would have achieved even half as much without you. Ben Jones, for kindly accommodating me when I came to use the ETC's workshop (and inevitably making one hell of a mess!). Ruth Morgan at the Department of Security and Crime Science, for introducing me to the weird and wonderful world of forensic sand grain examination.

My heartfelt appreciation also needs to go out to: Andrea Sella in Chemistry; Steve Etienne, Nick Constantino & Richard Thorogate in the LCN; Kevin Lee in EE; Piter, Tom & Trevor in the workshop; Andrew, Gerald, and their horde of helpful minions in the electronics lab; and all the other academic, support and administrative staff that make EE a great department. Thanks also to my friends and colleagues too numerous to list, both within UCL and without.

Finally, thank you to my partner and my family for their unwavering support and encouragement these past few years – I am forever grateful.

Contents

List of Figures **ix**

List of Tables **xiv**

List of Abbreviations **xv**

1 Introduction **1**

1.1 Objectives 2

1.1.1 Data recovery from damaged SIM cards 3

1.1.2 Examination of quartz sand grains 5

1.2 Outline 6

Bibliography 10

2 Atomic force microscopy **12**

2.1 Development of the atomic force microscope 13

2.1.1 Comparison of AFM to other forms of microscopy 15

2.2 Atomic force microscope components 20

2.3 Operating modes 23

2.3.1 Contact mode 25

2.3.2 Dynamic modes 29

2.4 Electric AFM techniques 37

2.4.1 Electric force microscopy 37

2.4.2 Scanning Kelvin probe microscopy 40

Bibliography . . . 46

3 SIM cards 48

3.1 History of smart cards . . . 48

3.2 SIM card structure . . . 52

3.2.1 SIM card contact pads . . . 54

3.3 SIM card microcontroller components . . . 56

3.4 Smart card memory . . . 58

3.4.1 Floating gate transistors . . . 59

3.4.2 Operation mechanisms . . . 60

3.4.3 NOR and NAND configuration . . . 72

3.4.4 Multi-level cell memories . . . 74

3.4.5 Memory reliability . . . 75

3.5 Fire investigation study . . . 83

3.6 Forensic evidence within SIM cards . . . 86

Bibliography . . . 89

4 Sample preparation 92

4.1 Early SIM card evidence analysis . . . 94

4.1.1 Energy dispersive X-ray spectroscopy . . . 94

4.1.2 Exhibit examination . . . 96

4.2 Resins: decapsulation . . . 100

4.2.1 Thermal/mechanical decapsulation . . . 105

4.2.2 Solvent decapsulation . . . 106

4.2.3 Fuming nitric acid decapsulation . . . 110

4.3 Resins: mounting . . . 112

4.4 Mechanical silicon removal . . . 119

4.5 Wet chemical etching . . . 127

4.5.1 Silicon dioxide removal . . . 127

4.5.2 Silicon removal . . . 130

4.5.3 Silicon removal: tetramethylammonium hydroxide 134
4.6 Summary of preparation process 148
Bibliography 152

5 Sample analysis – focused ion beam, Kelvin probe, and sample rewiring 158

5.1 Focused ion beam 158
5.1.1 FIB cross-sectional examination of SIM memory array . . 161
5.2 Applicability of Kelvin probe examination 163
5.2.1 Kelvin probe results 164
5.3 Rewiring / data retention 166
Bibliography 172

6 Microscopy results 173

6.1 Atomic force microscopy 173
6.2 Electric force microscopy 176
6.3 Scanning Kelvin probe microscopy 178
Bibliography 187

7 Quartz grain analysis 188

7.1 Introduction 188
7.2 Materials and methods 191
7.2.1 Instrumentation 191
7.2.2 Samples 191
7.3 Results 193
7.3.1 Topographical AFM scans 193
7.3.2 Statistical analysis 193
7.3.3 Applicability of lateral force microscopy 198
7.3.4 Fractal nature of grain surfaces 201
7.4 Discussion 205

7.4.1 Discriminant analysis theory 206
7.4.2 Discriminant analysis output explanation 210
7.4.3 Canonical discriminant functions from original sample sets 212
7.4.4 Final canonical discriminant function 213
7.5 Conclusions 221
Bibliography 224

8 Conclusions and future work 228

8.1 Conclusions 228
8.1.1 SIM card processing and data extraction 229
8.1.2 Quartz grain surface texture analysis 232
8.2 Future work and further applications 234
Bibliography 241

Appendix A Lapping/polishing results 243

Appendix B BOE etching results 245

Appendix C TMAH etching results 246

Appendix D Kelvin probe and baking results 248

List of Figures

2.1 AFM 'light-lever' laser detection system 22
2.2 Schematic of AFM cantilever and SEM micrographs of cantilever and tip 24
2.3 AFM force regimes 25
2.4 AFM contact mode force regime 26
2.5 Lateral force microscopy operation 29
2.6 Dynamic AFM signal modulation modes 31
2.7 AFM non-contact mode force regime 33
2.8 AFM intermittent-contact mode force regime 36
2.9 Lift-mode EFM operation schematic 38
2.10 CPD basic measurement setup 41

3.1 Global mobile subscriptions 51
3.2 Cross section of SIM chip module area 53
3.3 Smart card contact pads 55
3.4 Smart card contact pad designs 55
3.5 Von Neumann smart card architecture 56
3.6 Infineon SLE 66CX160S microcontroller 57
3.7 Floating gate transistor 59
3.8 FGT during HCI programming 62
3.9 HCI programming energy band diagram 63
3.10 'Lucky Electron' model 64

3.11 FGT during uniform and drain-side FNT programming 65
3.12 FNT programming energy band diagram 67
3.13 FNT erasing energy band diagram 69
3.14 FGT during uniform and drain-side FNT erasing 69
3.15 FGT cell I-V characteristics 71
3.16 FGT during reading 72
3.17 NOR and NAND architecture 73
3.18 MLC vs. SLC voltage distributions 75
3.19 Logic margin window vs. endurance 77
3.20 Data retention time vs. tunnel oxide thickness 80
3.21 Data retention graph for 100 year industry standard memory . . 83
3.22 House fire experiment temperatures 85

4.1 Photographs of broken SIM card presented for examination . . . 94
4.2 EDS principle 95
4.3 SEM image and EDS X-ray maps of broken SIM 97
4.4 EDS spectra for glob top and card body 99
4.5 SIM card silicon damage 100
4.6 Partially decapsulated SIM chip 108
4.7 Chip module inaccessible using solvents 109
4.8 Ideal die setting 113
4.9 Resin ratio tests 115
4.10 Bubbles congregating on die edges 116
4.11 Improved cast from magnetic stirring and ultrasonic agitation . . 117
4.12 Uneven die setting 117
4.13 Processing outcomes and delamination issues 119
4.14 Sample finish with various abrasives 121
4.15 AFM scan of P1200 surface finish 122
4.16 AFM scans of 1μm diamond surface finish 123

4.17 AFM scan of 6μm diamond surface finish and line profiles through scratches 123
4.18 Effect of etching various surface finishes 124
4.19 The effect of scratches on TMAH etching 125
4.20 The effect of scratches on choline hydroxide etching 125
4.21 Optical micrograph of selective scratch etching 126
4.22 Photolithographic mask for etch rate tests 129
4.23 BOE etch rate 130
4.24 Choline hydroxide 133
4.25 Tetramethylammonium hydroxide 135
4.26 Silicon hemisphere 137
4.27 Schematic of etching setup on hotplate 140
4.28 TMAH Si(100) etch rate vs. temperature, varied presence of agitation 141
4.29 Measured roughnesses for agitated and non-agitated TMAH solutions 142
4.30 Si(100) etch rate vs. temperature for agitated TMAH and TMAH & IPA solutions 143
4.31 Measured roughnesses for agitated TMAH and TMAH & IPA solutions 144
4.32 TMAH+IPA Si(100) etch rate vs. temperature, varied presence of agitation 145
4.33 Measured roughnesses for agitated and non-agitated TMAH & IPA solutions 147
4.34 Final sample preparation process 149

5.1 Gas-assisted FIB etching 159
5.2 Gas-assisted FIB deposition 160
5.3 FIB milled trench 161

5.4 SEM comparison of EEPROM/flash structure 162
5.5 Kelvin probe surface potential baking results 165
5.6 Kelvin probe change in surface potential results 166
5.7 SIM chips wire-bonded to chip carriers 168
5.8 Custom adapters 168
5.9 Other adapters used 168

6.1 Flash memory topography scan 174
6.2 Flash memory extracted topography profiles 175
6.3 EFM scan of flash memory array 177
6.4 SKPM scan of flash memory array 180
6.5 AFM tip comparison to trench size 181
6.6 Example of surface potential analysis 182
6.7 Comparison of topography and potential profiles 183
6.8 Gaussian-fitted histogram of tunnel oxide regions 184
6.9 Gaussian-fitted histogram of control regions 185

7.1 SEM images of grains from sample set 1 192
7.2 SEM images of grains from sample set 2 192
7.3 SEM images of grains from sample set 3 192
7.4 Sample set 1 AFM/amplitude scan 194
7.5 Sample set 1 AFM/amplitude scan offset from before, with 3D reconstruction 195
7.6 Sample set 2 AFM/amplitude scan showing sharp edges 196
7.7 Sample set 2 AFM/amplitude scan showing euhedral growths . . 196
7.8 Sample set 3 AFM scan showing upturned plates 196
7.9 Sample set 3 AFM scan showing rounding of edges 197
7.10 Example height distributions from sample set 1 and 3 199
7.11 Diamond AFM tip LFM damage 200

7.12 Fractal comparison image 1: 3D reconstruction, height distribution, and fractal dimension at 80µm^2 202
7.13 Fractal comparison image 2: 3D reconstruction, height distribution, and fractal dimension at 20µm^2 203
7.14 Fractal comparison image 3: 3D reconstruction, height distribution, and fractal dimension at 5µm^2 204
7.15 Canonical discriminant function plot for topography and derived height distribution statistics 214
7.16 Canonical discriminant function plot for fractal dimensions . . . 216
7.17 Canonical discriminant function plot for statistical roughness measures 218
7.18 Canonical discriminant function plot all statistical measures, with new data included 219

List of Tables

3.1 Smart card dimensions 52
3.2 EMC constituents 53
3.3 Retention time as a function of gate current 79
3.4 Data retention during house fire 86

4.1 Depot 1 constituents 106
4.2 Depot 2 constituents 107
4.3 Solvent soak decapsulation results 107
4.4 Lapping grit sizes 120
4.5 Polishing grit sizes 121
4.6 Silicon hemisphere etch rates 138

5.1 Rewiring success rates 170

6.1 Comparison of Gaussian distribution parameters 184

7.1 Fig. 7.5 contrasting region roughness values 194
7.2 CDF output tables for topography and derived height distribution statistics 213
7.3 CDF output tables for fractal dimensions 215
7.4 CDF output tables for statistical roughness measures 217
7.5 Final CDF output tables for all statistical measures, with new data included 220

List of Abbreviations

Below is an alphabetical list of commonly used abbreviations used throughout the report as a helpful reference to consult if necessary.

ABS	Acrylonitrile Butadiene Styrene
AC	Alternating Current
AFM	Atomic Force Microscopy
BOE	Buffered Oxide Etch
CDA	Canonical Discriminant Analysis (see DA)
CHE injection	Channel Hot Electron injection (see HCI)
CPD	Contact Potential Difference
DA	Discriminant Analysis
DC	Direct Current
DI water	De-Ionised water
EDS	Energy Dispersive X-ray Spectroscopy (also EDX)
EEPROM	Electrically Erasable Programmable Read-Only Memory (also E^2PROM)
EFM	Electric Force Microscopy
EMC	Epoxy Moulding Compound
FIB	Focused Ion Beam
FGT	Floating-Gate Transistor
FNT	Fowler-Nordheim Tunnelling
FSS	Forensic Science Service, Ltd.

GSM	Global System for Mobile Communications (originally Groupe Spécial Mobile)
HCI	Hot Carrier Injection
HEI	Hot Electron Injection (also known as HCI)
HF	Hydrofluoric Acid
IC	Integrated Circuit
IPA	Isopropyl Alcohol (Isopronanol)
IPD	Inter-Polysilicon Dielectric
ITRS	International Technology Roadmap for Semiconductors
KOH	Potassium Hydroxide
MLC	Multi-Level Cell
MOSFET	Metal-Oxide-Semiconductor Field-Effect Transistor
NAND	Logical “Not AND”
NOR	Logical “Not OR”
PVC	Polyvinyl Chloride
SEM	Scanning Electron Microscopy
SIM	Subscriber Identity Module
SKPM	Scanning Kelvin Probe Microscopy (also KPFM and SSPM)
SLC	Single-Level Cell
SMS	Short Message System (text message)
SPM	Scanning Probe Microscopy
STM	Scanning Tunnelling Microscopy
TMAH	Tetramethylammonium Hydroxide
UMTS	Universal Mobile Telecommunications System (3G alternative to GSM)
USB	Universal Serial Bus
USIM	Universal Subscriber Identity Module

Chapter 1

Introduction

The atomic force microscope (AFM) combines unparalleled non-destructive imaging capabilities with a flexibility allowing many types of samples to be imaged in various environments. Its lateral resolution is hundreds of times better than the diffraction limit of traditional optical systems thanks to highly accurate piezoelectric scanners. The microscope can perform standard height measurements by essentially 'feeling' the surface with an ultrasharp probe, producing a three dimensional image of the surface topography. In more advanced modes, the AFM can probe other interaction forces allowing various experiments to be conducted with mostly identical equipment. This tool has found a multitude of applications across a great many disciplines as diverse as museum curatorship, nano-biology, materials science, semiconductor fabrication, and has recently been finding adoption in the field of forensics.

Forensic science has always had a close link to science and engineering due to the rigorous, recreatable experiments and protocols that must be developed and adhered to for any forensic technique to be admissible as evidence. Following some recent critical government reports in the US [1] and the UK [2], as well as recent controversies that have plagued the field, *e.g.* bite mark analysis[1]

[1]www.nytimes.com/2007/01/28/weekinreview/28santos.html

and comparative bullet-lead analysis[2], there has been a push in recent years to provide greater validity and rigour to forensic protocols. With the exception of DNA analysis, many forensic science techniques were found to not offer a high enough degree of certainty in their conclusions and error margins, relying too heavily upon qualitative interpretation. In light of these criticisms, the AFM could become the perfect tool for nanometre-scale forensic examination, producing calibrated, quantitative, and repeatable results.

1.1 Objectives

Two distinct projects were undertaken for this PhD, both dealing with a forensic application of the AFM, and both dealing with samples that are commonplace the world over – sand grains and SIM cards. It is important to state up front that due to the greater time, cost, and expertise required, the SIM card process developed is primarily for use in high-value cases on damaged equipment. The recovery of partial SIM cards or immolated handsets during terrorism or organised crime investigations could present forensic examiners with devices damaged beyond simple rewiring, requiring deep binary extraction of data directly from the memory array. The quartz sand grain examination, however, offers a relatively high-volume technique due to the possibility of automation and a much greater throughput.

In addition to the work described within this transcript, a placement was carried out between May–August 2010 at The Forensic Science Service[3] in their Electronic Forensic Science Unit (EFSU) based at their Lambeth headquarters. The aim of this placement was to conduct a feasibility study into the possible development of a more generic mobile phone flash memory data extraction tool using Joint Test Action Group (JTAG) to read the built-in flash memory modules

[2] www.fbi.gov/news/pressrel/press-releases/fbi-laboratory-announces-discontinuation-of-bullet-lead-examinations

[3] The Forensic Science Service® was a trading name of Forensic Science Service Ltd., a UK Government owned company.

in situ (on a mobile phone motherboard). This was designed to address the issue of those few mobile phone handset models that could not be interrogated using their existing in-house tools.

1.1.1 Data recovery from damaged SIM cards

The FSS identified a need for a new data extraction tool to be developed for use in their EFSU in cases where criminal evidence involved mobile telecommunications equipment that was damaged, resulting in the failure of traditional data extraction methods. The primary PhD project focused on the development of this forensically-sound process for the recovery of data from non-volatile memory, with specific focus on using the AFM to examine mobile phone SIM cards.

SIM cards, present in every mobile phone and other similar devices, *e.g.* tablets, hold a variety of data of particular value to forensic examiners. This information can provide a critical link between a suspect and recorded actions, whereabouts, contacts, messages, etc. Broken SIM cards present an obstacle to forensic examiners; damage to contact pads or bond wires require the microcontroller to be removed and rewired for electronic interrogation. However, any damage to the microchip leaves examiners unable to perform standard electronic interrogation, and leaves valuable data beyond their reach.

The data stored within a SIM card's non-volatile memory array takes the form of charged or discharged electrically isolated structures within floating gate transistors. The amount of charge held within these 'floating gates' determine the level of screening on the normal function of the transistor – either allowing or inhibiting the conductive source-drain channel formation when a gate voltage is applied. Thus, the level of charge residing in the floating gate ultimately determines the logical state of the transistor. Typically, EEPROM[4] and flash memories have an incredible degree of longevity and robustness – able to store

[4]EEPROM stands for Electrically Erasable Programmable Read-Only Memory (also known as E^2PROM)

data for upwards of 100 years, and withstand the sorts of temperatures commonly used by criminals to destroy evidence.

Approaching from the topside of the chip, a forensic examiner would be faced with removing a large quantity of unknown structures and layers, of various materials and dimensions, before any direct probing of the microcontroller's memory array could be attempted. Many of these structures would have to be individually and selectively removed to avoid damaging the protective oxide layer surrounding the floating gates. Given the difficulties presented by this approach, in part due to the amount of unknown parameters, this topside approach is inadvisable. However, approaching from the opposite side would reduce the number of obstacles to a single homogeneous structure – the bulk silicon substrate upon which the chip was built. It is this forensically-sound back-side approach which has been developed and is described in this thesis.

A similar approach had previously been tested for feasibility by a group working with the French space agency CNES, published between 2005 and 2006 [3–5]. Working with samples whose final application was aerospace and astronautics programmes, their samples were decidedly not state-of-the-art technology. As is customary within these fields, the most up-to-date technology is not typically applied given its lack of rigorous testing and proven reliability. As such, the fabrication process node that their sample devices were manufactured to had been adopted approximately ten years prior to the date of their publications.

The field of data recovery has been very popular for many years, both in the commercial realm and research fields of computer science and microscopy. Due to market forces, these tended to focus on magnetic hard disk drive technology, or if solid-state devices – electronic recovery methods only. To our knowledge the only study carried out into the area of data recovery from electronic memory devices using scanning probe microscopy techniques are those by De Nardi *et al.* mentioned above.

These publications laid the ground work for Jones[5] & Kenyon[6] to conduct a feasibility study [6,7] into the possible application of this method to retrieve data from damaged smart cards. Smart cards were programmed with sample data via a PC interface, and heated and held at three elevated temperatures. The chips were rewired and attempts were made to read the data through the same PC interface. Those heated to 180°C showed no loss of data, but those heated to 450°C and above proved faulty, suffering heat-induced material damage. One particularly notable observation covers the variety of encapsulants encountered in chip packaging, and the statement that different protocols must be developed to extract the chips for any further developed technique.

It is from this basis that the primary investigation detailed in this thesis began – to build upon these concepts, develop a viable forensic extraction and preparation process, and assess the use of AFM to determine if the methods are still applicable to modern devices. The research followed a linear path early on due to the prerequisite of having a well developed and reliable sample preparation process. With this completed, it was possible to assess the application of the AFM to directly probe the exposed memory arrays, and to ascertain the SIM chip reliability & robustness, further clarifying the need for such an extensive process. Issues such as reliability and sample preservation had to be considered throughout the project to develop this as a forensically-sound technique.

1.1.2 Examination of quartz sand grains

Examination of sand grains can enable discrimination between samples pertinent to a forensic investigation. Many analytical techniques exist in distinguishing grain provenance, and the analysis of the morphological textures on grain surfaces has been demonstrated to be able to provide highly discriminatory results. The traditional SEM examination of quartz sand grain micro-textures is primar-

[5]Experimental Techniques Centre, Brunel University
[6]Department of Electronic & Electrical Engineering, University College London

ily a qualitative technique, relying upon skilled users' discriminatory abilities rather than numerical figures of merit to distinguish between grain provenance.

The second project undertaken for this PhD covered the potential application and development of AFM as a complimentary quantitative technique to SEM examination of quartz grain surface textures for forensic analysis. Numerous AFM topographical images of quartz sand grains of different known origins and history were taken. From these scans, various statistical figures of merit were computed and this data was analysed using a statistical method known as canonical discriminant analysis to generate multivariate discriminant functions. These functions assess and weight the individual measures by their discriminatory ability, and if significant, can be further used to predict the classification of unknown samples.

1.2 Outline

Chapter 2 covers the background theory behind the AFM, starting with a short explanation of the history and comparison with other microscopy techniques. This is followed by a brief description of the component equipment and a look at the theory behind different AFM operating modes. Finally, there is a closer look at two advanced electrical scanning probe techniques pertinent to the work outlined in this thesis.

Chapter 3 describes the SIM smart card from various perspectives. Beginning with the history of smart cards, the chapter will then proceed to outline the structure of a SIM card, followed by the components typically found in cryptographic smart card microcontroller circuits, then focusing in particular on the EEPROM/flash memory structure and operational mechanisms such as hot carrier injection and Fowler-Nordheim tunnelling. This is followed by an overview of NOR and NAND architecture, and a look at the multiple-level cell capabilities of modern flash memories. The reliability and endurance is then discussed, with

specific focus on the retention times and accelerated testing methodologies used in military/industrial applications – this is contrasted with the data gathered from a controlled house fire investigation. Finally, the chapter is concluded with an overview of the potential forensic evidence held within mobile phone SIM cards.

Chapter 4 begins with the analysis of the card structure using electron microscopy and X-ray spectroscopy, and failed early attempts at device extraction from this obtained SIM evidence. Thereafter follows the development of the sample preparation process that is fundamental to the SIM card data recovery project. Beginning with the development of an effective decapsulation technique, numerous methods were tested on a multitude of devices and documented accordingly. The extracted chips must then be re-encapsulated within an epoxy resin cast for further silicon removal processing, common problems and solutions are again documented. The first silicon removal stage involves mechanical lapping and polishing – proven methodology is outlined for this process and advice offered as to the ideal remaining silicon thickness and final surface finish requirements. The second silicon removal process involves wet chemical etching of the remaining silicon using an organic-ballasted ammonium hydroxide solution – this process is detailed and characterised at some depth, with advice given to obtain optimum etching uniformity, thus improving sample preservation.

Following on from the decapsulation method developed for SIM cards, Chapter 5 details various experiments conducted to better analyse SIM card embedded memories. Starting with a brief working theory of the focused ion beam, examination of a SIM memory array was conducted by repeated sectioning, with selected cross-sections compared to previously documented EEPROM devices. The second and third experiments both make use of an accelerated testing method known in the field of failure analysis as a 'stabilisation bake'. With samples pre-programmed with sample data and then exposed to greatly elevated temperatures for set periods, the applicability of a scanning Kelvin probe (not

to be confused with the AFM-based technique) for initial sample examination is given. Finally, a data retention study is conducted using these stabilisation-baked samples – mounting and rewiring using a wedge wire bonder, reading is attempted to assess the data retention characteristics after exposure to elevated temperatures.

The final SIM card work is covered in Chapter 6, which covers the application of two related electrical scanning probe microscopy techniques – electric force microscopy (EFM) and scanning Kelvin probe microscopy (SKPM). These techniques were conducted on samples prepared according to the earlier developed processing methodology, and results are shown for AFM, EFM and SKPM. Statistical distributions of floating gate potentials are constructed and discussed, along with a discussion on the limitations of the techniques.

Chapter 7 covers the second project undertaken for this PhD – the examination of quartz sand grains for forensic analysis. The chapter starts with a discussion outlining current quartz grain examination methodology along with its limitations and shortcomings, recently documented forensic applications of the AFM, and why the AFM has the potential to supplement traditional quartz grain forensic analysis. The initial sample sets are then shown in SEM micrographs, and surface micro-topography examined using an AFM, topographical statistical measures and height distributions are then taken and compared. This is followed by a short discussion on the applicability and dangers of lateral force microscopy, and a look at the statistically self-similar fractal nature of grain surfaces across different scales.

In the discussion section, the background theory and explanation of the statistical output is given for canonical discriminant analysis, a technique used to ascertain the discriminatory ability of the various statistical measures obtained. The initial three sample sets are then analysed using this method, focusing on three related groups of statistical measures. New data from the initial samples and a fourth sample set, is then included and compiled over all figures of merit

into a single set of computed multivariate discriminant functions. Finally, the project is concluded with a discussion of the limitations of the study, the potential application of the AFM as a complimentary technique to traditional SEM analysis, and an overview of the potential automation possible with this tool.

Finally, Chapter 8 contains a summary of the main conclusions of this thesis and application-specific contributions to the fields of microscopy and forensics, along with a discussion of the main obstacles to be overcome in fully realising the development of these methods, and ideas for further research areas to extend this work in the future.

Bibliography

[1] National Research Council of the National Academies, *Strengthening Forensic Science in the United States: A Path Forward*. The National Academies Press, 2009. [Online]. Available: http://www.nap.edu/catalog/12589.html

[2] The Law Commission, *Consultation Paper No 190: The admissibility of expert evidence in criminal proceedings in England and Wales - A new approach to the determination of evidentiary reliability*. The Law Commission, 2010, no. 190. [Online]. Available: http://www.justice.gov.uk/lawcommission/docs/cp190_Expert_Evidence_Consultation.pdf

[3] C. De Nardi, R. Desplats, P. Perdu, F. Beaudoin, and J. Gauffier, "Oxide charge measurements in EEPROM devices," *Microelectronics and Reliability*, vol. 45, no. 9-11, pp. 1514–1519, 2005.

[4] ——, "EEPROM Failure Analysis Methodology - Can Programmed Charges Be Measured Directly by Electrical Techniques of Scanning Probe Microscopy?" in *31st ISTFA 2005*. San Jose, CA: ASM International, 2005, pp. 256–261.

[5] C. De Nardi, R. Desplats, P. Perdu, J. Gauffier, and C. Guerin, "Descrambling and data reading techniques for flash-EEPROM memories. Application to smart cards," *Microelectronics and Reliability*, vol. 46, no. 9-11, pp. 1569–1574, 2006.

[6] B. J. Jones, "Burnt to memory - data extraction from heat damaged mobile phones," *Public Service Review: Home Office*, no. 15, pp. 68–70, March 2007. [Online]. Available: http://bura.brunel.ac.uk/bitstream/2438/695/1/PSR_HO_Jones_15_2007_68.pdf

[7] B. Jones and A. Kenyon, "Retention of data in heat-damaged SIM cards and potential recovery methods." *Forensic Science International*, vol. 177, no. 1, pp. 42–6, May 2008.

Chapter 2

Atomic force microscopy

Atomic force microscopy (AFM) offers a truly unrivalled range of techniques for examining a multitude of samples down to the nanometre scale. This chapter gives an overview of the historical development of the AFM, a brief comparison to other microscopy techniques, an outline of the experimental systems, and details of specific AFM techniques (and variations thereof) relevant to the work featured in this transcript – the methods covered in this chapter are far from exhaustive given the huge variety of experiments available with an AFM. The AFM is fundamentally different from other microscopy techniques. Rather than using focused light or electrons to create an image, it raster scans a sharp probe across the sample, sensing the physical interactions between tip and sample as it does so. From this, the AFM constructs a map of sample topography.

The AFM is part of a group of techniques described by the term scanning probe microscopy (SPM), a branch of modern microscopy techniques that use very sharp probes to analyse the surface of a sample, sometimes down to atomic scale resolution. Various interactions between the probe tip and sample surface can be measured as the probe is scanned across the sample surface, building up a map of the surface with respect to the specific force being detected.

2.1 Development of the atomic force microscope

An early predecessor to modern SPM techniques is the stylus profiler, invented by Schmalz in 1929. This machine essentially dragged a very sharp needle across the surface of a sample to measure feature heights, an imaging mode we would refer to now as 'contact mode'. One main drawback of this method was that the probe, if harder than the sample, could damage the sample surface if it collided with large features. Newer 'profilometers' have advanced greatly from these early models and are still in use in numerous applications worldwide.

Young, Ward, and Scire published their paper on the 'topografiner' in 1972 [1]. This microscope advanced upon the stylus profiler by using the dependent relationship between the tip-sample distance and the electron field emission current (for a conductive sample) to maintain a close proximity to the surface. It also solved much of the problem of impact damage caused by the stylus through the use of piezoelectric elements to control probe location and tip oscillation.

In 1981 Binnig and Rohrer, working at IBM, developed further upon Young's topografiner by increasing the vibrational isolation of the microscope and improving feedback control. This was the invention of the scanning tunnelling microscope (STM). With this new feedback control system they were able to monitor tunnelling current between a conductive sample and the sharp probe tip. Electron tunnelling has a far greater dependence to the tip-sample distance than field emission; so great is this dependence that effectively only the atom at the tip of the sharp probe is interacting with the sample surface through tunnelling. This fact actually helped in the production of suitable tips, since only the very tip is important and structure further along can be ignored. Binnig and Rohrer shared half of a Nobel Prize in Physics in 1986 for their development of the STM – that such a small device could image atoms and atomic structure that could otherwise only be discerned by transmission electron microscopy and X-ray diffraction patterns was incredible. The STM, however, had one main

drawback: it could only image electrically conductive samples.

In 1986 Binnig, Quate and Gerber published a paper entitled 'Atomic Force Microscopy' [2]. In this paper they describe how they overcame the limitation of the STM to image only conductive samples. Modifying the STM, they replaced the wire tip with a thin gold foil strip (the cantilever) upon which they had mounted a tiny diamond shard (the tip). The vertical movement of the cantilever as it is scanned across the surface of a sample was monitored by measuring the tunnelling current between the gold strip and an STM probe suspended above the cantilever – this was the first AFM. To quote their paper:

> "*The atomic force microscope is a combination of the principles of the scanning tunnelling microscope and the stylus profilometer*"

One advancement on the original AFM was developed by Martin *et al.* in 1987 [3]. An effective simplification of the previously complex STM detection method, this advancement used a vibrating cantilever and a light-lever technique. This light-lever was first developed by Schmalz in 1929 to amplify the distance of movement from a surface profiler.

Modern AFMs use microfabricated silicon cantilevers which can have a large variety of configurations, shapes, materials, platings, etc. to suit the type of sample or force to be monitored. In addition, they use piezoelectric scanners for high spatial resolution, and their detection systems, while based on the original light-lever used by Schmalz and Martin *et al.* , are greatly modernised, using a laser reflected off the back of the cantilever onto a position-sensitive photodiode to monitor deflection.

The AFM, and subsequent SPM techniques built upon this foundation, can measure a multitude of tip-sample interactions such as: van der Waals forces, capillary forces, capacitance, magnetism, coulombic interaction, tunnelling current, chemical bonding, etc. They also have the capability of measuring a broad range of samples in a variety of media, such as ambient atmosphere, vacuum and

liquids. It is also capable of imaging samples over a large range of temperatures (from 10K to 500K). It is this versatility that has guaranteed AFM a place at the forefront of scientific research today.

2.1.1 Comparison of AFM to other forms of microscopy

An AFM is a very different form of microscopy from traditional optical microscopes; they both require little or no sample preparation, but the data output from an AFM in the standard topography scans (sometimes referred to as colour maps) forms an array of sampled points built up by scanning the tip across the surface in a raster scan pattern. This is very different to conventional microscopy techniques, such as optical microscopy where reflected light from the sample surface is focused by an array of lenses and viewed by the user or captured in a micrograph image. Imaging takes far longer with an AFM, but provides unambiguous height measurements regardless of the reflectivity of surfaces. Many AFMs incorporate an optical microscope into the imaging suite, allowing a user, once zeroed in to the tip, to locate areas they wish to scan quickly and easily.

The AFM offers the user a method of imaging samples in a range of media and environmental conditions, *e.g.* biological samples *in situ*. With some additional equipment a range of tip-sample interactions can be detected, allowing the basic AFM setup to be used to conduct a wide variety of SPM experiments. Given that AFM was built upon the basic principles of STM, the two techniques share many common components and design principles with one another and with other SPM techniques. The identical piezoelectric scanning equipment used means that both AFM and STM can achieve a similar resolution and offer a versatile range of scanning methods – as well as taking raster scans over a target area, the slow scan axis can be disabled to take single line scans in a specified direction, and can even pinpoint sections of a previous scan to conduct point spectroscopy measurements.

Both AFM and STM, unlike other forms of traditional microscopy, have no

depth of field or focus, and require no illumination to obtain an image. The data gathered is digital, pre-calibrated (for many sample interactions), and inherently three dimensional. Such digital data sets allow post-imaging data manipulation, *e.g.* cropping, the application of plane and background correction algorithms and data filters, and the extraction of cross sections at any angle through the data to investigate the geometry/surface readings of a feature. It also lends itself to statistical analysis of the data, with various roughness measures commonly used to analyse topography, as well as more advanced methods, *e.g.* height data histograms, Fourier analysis, and fractal dimensions.

One of the main advantages of AFM over optical microscopy is a greater spatial resolution, both laterally and vertically. The lateral resolution of an AFM system is typically around 1nm, with a vertical resolution around 0.1nm. The minimum resolvable lateral distance for an optical system, such as a microscope, is traditionally determined by the limit of its angular resolution, of the form:

$$d_{x,y} = \frac{A\lambda}{\mathrm{NA}} = \frac{A\lambda}{n\sin\theta} \tag{2.1}$$

Where: $d_{x,y}$ is the lateral (x, y) resolution limit, A is the coefficient corresponding to a specific limit, λ is the wavelength of light, NA is the numerical aperture of the objective lens, n is the refractive index of the medium, and θ is the half-angle of the maximum cone of light that can enter/exit the lens. The only ways to improve the resolution are to decrease the imaging wavelength, increase the numerical aperture, or image within a medium with a higher refractive index. Various values of the coefficient, A, can be used to represent different limits.

This limit on the angular resolution was discovered by Ernst Abbe in 1873, and defines the finite diameter of the disc/spot created by a point source; the coefficient of the Abbe diffraction limit is A=0.5. This was later refined by Lord Rayleigh in 1896 to distinguish the separation of two Airy patterns; the

Rayleigh criterion uses a coefficient of A=0.61. It is also possible to decrease the coefficient beyond the Abbe limit to A=0.47, the Sparrow limit; this defines the distance between two spots where there is no longer a decrease in the intensity between central peaks, but a constant brightness across the central region, while maintaining a distinction between both sources because of the extended image.

A commonly used approximation for the lateral resolution is $d_{x,y} = \lambda/2$. For red light (λ=650nm), this results in a lateral resolution of approximately 325nm for imaging in air. It is possible to improve the resolution of optical systems by increasing the numerical aperture, changing to a media with a higher refractive index, and using light of a lower wavelength. For example, by immersing the system into oil the numerical aperture can be increased to 1.4; and by using green light (λ=550nm) the resolution achievable is just under 200nm. Other forms of electromagnetic radiation, with wavelengths below the visible spectrum can be used in specialised microscopes, *e.g.* UV (λ=100–400nm) microscopes, to further surpass the resolution limit.

For optical microscopy, the axial (z-axis) resolution, d_z, is of the form:

$$d_z = \frac{2\lambda}{\mathrm{NA}^2} = \frac{2\lambda}{(n \sin\theta)^2} \tag{2.2}$$

In practice, this results in a z-resolution 2–3$\times$ worse than the lateral resolution; typically above 500nm. Thus, optical microscopy clearly cannot compete with the digitised, sub-Å resolution, 3D height data gathered by an AFM. It is common to compare an AFM to a mechanical stylus profilometer, specifically when discussing the vertical resolution and tip forces exerted on a sample. The profilometer uses a sharp metal/diamond probe to scan across a sample surface with a predefined contact force, mapping out the sample topography and allowing the step height and/or roughness of a surface to be measured.

The contact force exerted on a sample surface by a profiler tip is typically of the order of 10^{-6}N, substantially larger than an AFM, which exerts a force

of around 10^{-9}N on the sample when operating in tapping mode, making AFM a far less destructive technique. Lacking the more complex feedback systems of an AFM, common problems include bending the sharp probe and causing damage to the sample surface due to larger lateral tip forces. Many profilometers can measure variations in height <5nm, and can measure step heights up to many hundreds of micrometres, whereas an AFM, due to the sensitivity of the piezoelectric control elements, can attain a vertical resolution greater than 0.1Å, but is usually limited to a maximum height difference of 5–10μm.

The AFM is also often compared to electron beam microscopy techniques. Transmission electron microscopy (TEM), the original form of electron microscopy, is a technique whereby an electron beam is accelerated by high energy through a semitransparent sample, typically produced by thinning a sample to below 300nm thickness. The sample partially scatters the beam and upon emerging it carries information about the structure of the sample.

Scanning electron microscopy (SEM) images a sample differently from TEM, and more akin a traditional optical microscope; instead of passing an electron beam through a thinned sample, an electron beam is raster scanned across a sample surface. SEM imaging requires a sample to be either metallic or semi-conductive, with non-metallic samples requiring a deposited surface coating of gold/carbon prior to imaging, however, environmental SEM does not require this sample preparation step, making it particularly suited to biological samples. The lost beam energy at each point is converted into a variety of forms depending on composition and incident energy of the beam, such as: the reflection of back-scattered electrons, or the emission of light, X-rays, or low-energy secondary electrons.

Not all of these emission types are usually detected in a single machine; instead most tend towards the detection of low-energy (<50eV) secondary electrons ejected from inner shells by inelastic scattering interactions. To focus the electron beam, two or more magnetic lenses are used, the upper lenses being

condensers and the final lens the objective or 'probe-forming' lens. An SEM typically offers magnification between 10 and 500,000 times. The resolution of an SEM depends upon the attainable spot size of the electron beam, which is characterised by the wavelength of the electrons and the focusing system, but also by the specific sample material interaction with the beam. The wavelength of an electron, λ_e, often called the de Broglie wavelength, is governed by its momentum and shown by the de Broglie equation:

$$\lambda_e = \frac{h}{p} = \frac{h}{m_0 v} \tag{2.3}$$

Where: h is Planck's constant; and p is the relativistic momentum of the electron; m_0 is the rest mass of an electron; and v is the velocity. Accelerating the electron in an electric potential, U, gives a kinetic energy of eU; rearranging the kinetic energy gives:

$$v = \sqrt{\frac{2eU}{m_0}} \tag{2.4}$$

Substituting this back into Eq. (2.3) gives an electron wavelength of:

$$\lambda_e = \frac{h}{\sqrt{2m_0 eU}} \tag{2.5}$$

However, an SEM accelerates electrons to relativistic speeds by using voltages of several thousand volts, thus the momentum must be modified using the relativistic energy relation to take the form:

$$\lambda_e = \frac{h}{\sqrt{2m_0 eU}} \frac{1}{\sqrt{1 + \frac{eU}{2m_0 c^2}}} \tag{2.6}$$

Where: c is the speed of light. In this form the first half is the non-relativistic component, and the second half is the relativistic correction factor. Despite wavelengths of 2.5pm for a 200kV TEM, and 12.2pm for a 10kV SEM, the

resolution achievable depends greatly on the instrument. The actual achievable resolution can be between <1–20nm for an SEM, and <0.5Å for a TEM with modern aberration correction advancements.

2.2 Atomic force microscope components

A standard AFM consists of three core components: microscope stage, control electronics, and computer interface. The microscope stage is what most users would refer to as the microscope itself; it holds the probe and takes measurements of the sample. The control electronics handle the input signal generation, output signal analysis, proportional-integral-derivative (PID) bias control, digitising of output signals for examination, and feedback control systems to accurately control the movement of the piezoelectric elements and stepper motors. The computer interface is often used with AFM manufacturers' proprietary software suites to control the input to the stage and visualise and analyse the output data.

Since the AFM data is digital, already calibrated and specifically three-dimensional, many software suites allow for more than just simple image visualisation, containing a multitude of filters, analysis tools and imaging features. These include: various colour palettes and shading options, pseudo-3D visualisation of topography maps, algorithms to level the data using three-point plane fitting or 1st/2nd/3rd-order polynomials, high-pass filters to enhance edges, low-pass filters to eliminate high-frequency noise, multidimensional statistical calculations, image rotation and cropping, and user-defined cross-section line profiles, to name but a few.

The microscope stage contains numerous important features crucial to AFM operation. The stage itself must be mechanically rigid to minimise any vibration between a sample held on the sample holder and the probe imaging its surface. To isolate the AFM from external vibrations, the entire stage is often attached to

a floating isolation table and/or held within a sound-proof chamber. To rapidly locate and aid in positioning the tip over specific sample features an optical microscope is often built into the microscope. Given the computer interface used to control the AFM, it is common to have a digital video camera allowing positioning from within the software suite.

The scan head in an AFM is typically moved coarsely over the sample by x/y-stepper motors. Once a location has been found, fine lateral movement is achieved using an arrangement of piezoelectric elements. Piezoelectric materials are crystalline, ceramic or polymeric materials which convert electrical potential into mechanical motion. When a potential is applied across opposite sides of the material, it changes geometry – the magnitude depending on the geometry of the device and of the applied potential. The piezoelectric elements used in AFM scanners are typically constructed from amorphous lead zirconate titanate (PZT), $Pb[Zr_xTi_{1-x}]O_3$, $0 < x < 1$. They are most commonly configured in a tube shaped arrangement, allowing lateral motion by bending the tube in the x/y direction, and z-motion by elongating it. With tube scanners, it is important to accurately calibrate them beforehand, as this configuration exhibits high levels of non-linearity.

Vertical movement is controlled by both a step motor, which typically has a range of a few cm and a resolution of a few μm, and a z-piezo with an expansion range of <10μm and a resolution down to sub-angström distances. Using the stepper motor alone to approach the surface is dangerous given the large step size. The small step size of the piezoelectric elements is required, however, their small z-expansion range makes them insufficient to the task.

Combining both methods of z-motion, the AFM uses a 'woodpecker' technique to safely approach the surface. This involves carefully extending the z-piezo a distance, *e.g.* 5μm, to see if the tip encounters the surface. The piezo is retracted, and the stepper motor is extended a shorter distance, *e.g.* 1μm. This process repeats until the surface is located safely, then the feedback con-

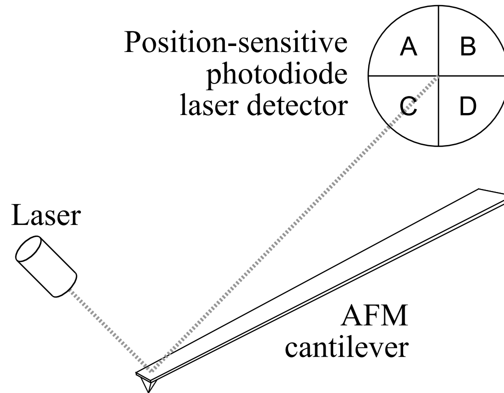


Figure 2.1: Simplified representation of the 'light-lever' laser detection system used to monitor deflection in the AFM cantilever. Note that the position detector is typically a quad-cell setup enabling the twist in the cantilever to also be monitored.

trol is engaged. This method allows the surface to be engaged from a large initial distance without crashing the tip into the sample, using the accuracy and 'gentleness' that only piezoelectric motion provides.

Once the surface has been safely approached, the tip is scanned in a raster pattern using the x/y piezoelectric elements to control fine movement. The term 'fast scan axis' is taken to mean the axis of the individual raster line motion. In regular AFM imaging, this is typically conducted in the direction approximately following the cantilever's (axial) beam. The so-called 'slow scan axis' is the slower progressive movement along the sample surface as each raster line is completed; this is perpendicular to the fast scan axis.

The forces between tip and sample are not measured directly, rather they are calculated using the deflection of the cantilever as it scans the tip across the surface. The most common detector used in modern AFMs is the light-lever configuration [4], see Fig. 2.1. While relatively simple to implement, this detector is highly sensitive to small cantilever deflections. By reflecting a laser off the top of the cantilever onto a position sensitive photodiode located a large distance away, small cantilever deflections result in large changes in spot position.

The feedback system passes the signal received by the photodiode back to

the piezoelectric scanner to maintain a constant tip-sample force, and thus a constant tip-sample distance. By monitoring the voltage applied to the piezoelectric scanner as the tip is raster scanned over the surface, a height map is constructed.

The basic setup of a modern AFM uses a sharp, nanoscale tip mounted on one end of a flexible microfabricated silicon cantilever, see Fig. 2.2. The cantilevers come in a range of shapes and sizes suitable for various operating modes. They are generally specified by their physical dimensions (width, length, thickness) and material (typically Si or Si_3N_4), which determine other important parameters (force constants and resonant frequency). Longer cantilevers with a much lower spring constant and a wider range of motion are ideal for contact mode imaging, while shorter cantilevers have a higher spring constant, thus a higher resonant frequency, making them suitable for oscillating modes. It is common to coat the top of the cantilever with a reflective Al/Au coating to boost laser reflection, this is known as a 'reflex' coating.

Other coatings are possible on both sides of the cantilever, including the tip, making them suitable for more advanced AFM modes, *e.g.* enabling electric or magnetic field detection. Some examples of more exotic probes are: all-diamond tips ideal for imaging hard, rough surfaces without tip damage; super-sharp tips for enhanced imaging; carbon nanotube AFM probes ideal for measuring high-aspect ratio features; or a 'hammer-head' design with tips laterally offset to one side of a hammer-head cantilever, designed to conduct force spectroscopy from the twist of the cantilever for every 'tap' in tapping mode (see Veeco Dimension Edge system with HarmoniX mode [5]).

2.3 Operating modes

Fundamentally, the AFM is a tool for examining the topography of a sample. The interaction forces between the tip and sample dictate the operating regime

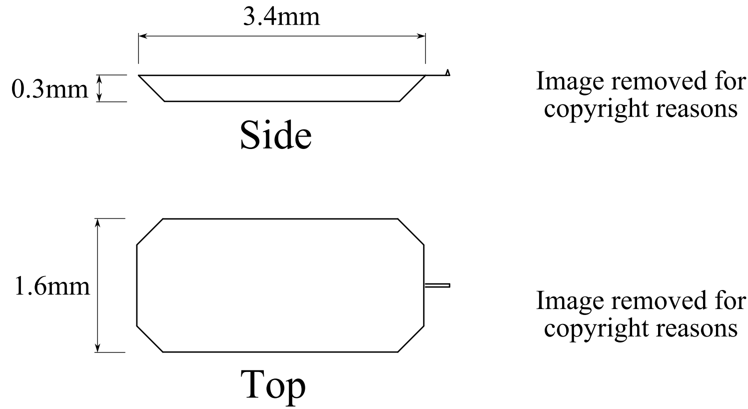


Figure 2.2: A schematic of the standard format chip carrying an AFM cantilever, along with scanning electron micrographs of a cantilever and close-up of an AFM tip. From: Bruker Corporation (www.brukerafmprobes.com).

and vary according to the distance from the surface, see Fig. 2.3.

The force/deflection-distance curve can be represented by the empirical equation of the Lennard-Jones potential (V_{LJ}), which approximates the interaction between a pair of neutral, unbonded atoms, as shown in Eq. (2.7). The r^{-12} term is the repulsive term, the r^{-6} term is the attractive term.

$$V_{\mathrm{LJ}} = 4\epsilon \left[\left(\frac{\sigma}{r} \right)^{12} - \left(\frac{\sigma}{r} \right)^{6} \right] \tag{2.7}$$

Where: ϵ is the depth of the potential well; σ is the atom diameter; and r is the interatomic distance. Some also represent AFM tip-sample interaction with the Buckingham potential (V_{B}), a simplification of the Lennard-Jones potential, see Eq. (2.8).

$$V_{\mathrm{B}} = Ae^{-Br} - \frac{C}{r^{6}} \tag{2.8}$$

Where: A, B, and C are constants; and r is the interatomic distance, as previously defined. These models, while an oversimplification of the tip-sample

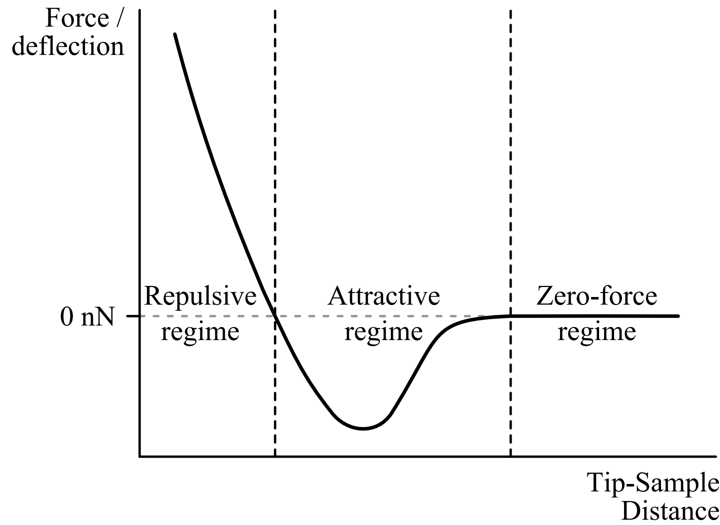


Figure 2.3: Force/deflection diagram showing the attractive, repulsive, and 'zero-force' regimes acting upon the AFM tip at varying tip-sample distances.

van der Waals interaction forces, offer an insight into the different regimes an AFM can operate in based on various tip-sample distances.

2.3.1 Contact mode

The immense popularity of AFM since its invention stems from the inherent accuracy and resolution attainable, but also its versatility as a basis for other SPM techniques, measuring numerous tip-sample interaction forces. Contact mode was the earliest form of AFM operation, and one of the simplest to understand. The tip assembly is carefully moved towards the sample. The cantilever deflection is proportional to the tip-sample interaction force, and this far from the surface these forces are practically non-existent. The tip will undergo an initial 'snap-in' towards the surface as it is advanced through the attractive interaction regime. As the tip continues to advance, the tip-sample interaction enters the repulsive regime, and the cantilever bends away from the surface accordingly, see Fig. 2.4.

The strong repulsive van der Waals interaction acting on the tip is a result of the exchange interactions from the Pauli exclusion principle, and acts at a range

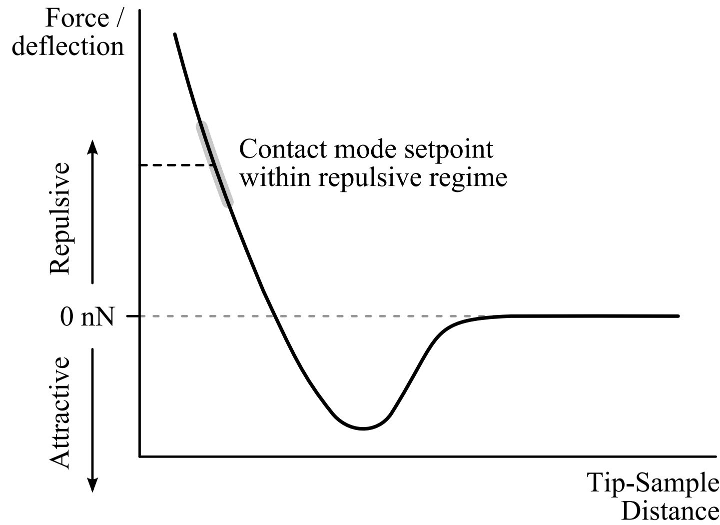


Figure 2.4: Force/deflection diagram showing the repulsive force regime acting upon the AFM tip during contact mode operation, with deflection setpoint shown.

of a few angströms. Contact mode techniques are able to image relatively rough samples, and offer higher scan speeds due to the direct relationship between cantilever deflection and topography, without the need for oscillation measurements and calculations required in other imaging modes.

As the tip is scanned over the sample, higher surface features will increase the cantilever deflection, while lower features reduce it. By monitoring this deflection, *e.g.* using a light-lever sensor incorporating a position-sensitive photodiode, a feedback system can adjust the tip assembly height accordingly by controlling the z-motion piezoelectric elements. The original cantilever deflection, and therefore the tip-sample interaction force, is maintained at a user-defined level throughout the scan using this feedback system. By recording the vertical adjustment of the tip assembly, a map of the surface topography is generated. This is known as constant-force contact mode AFM. The limitation of maintaining constant deflection lies in the ability of the feedback circuit to accurately and rapidly adjust the vertical movement of the cantilever assembly.

The force applied to the surface by the tip is described by Hooke's Law: $F = -kx$. Where: F is the (restoring) normal tip-sample force; k is the can-

tilever's spring constant (stiffness); and x is the deflection distance. The feedback system maintains the deflection distance of the cantilever at an operator-defined constant setpoint. A greater force will be applied to the sample by using a higher spring constant cantilever, or maintaining a greater deflection setpoint.

The primary disadvantage of contact mode AFM is the high lateral forces generated as a result of the large normal tip-sample force, and the constantly maintained contact between the tip and sample surface during scanning. This can distort images and deform or destroy soft samples, or remove weakly adsorbed sample material. As such, contact mode AFM is generally restricted to harder, more robust samples.

Many samples will experience an additional force because of the presence of a thin capillary layer of water present on the sample (and tip) surface. This force, often dwarfing the force applied due to the setpoint value, acts to pull the tip towards the surface, generating large forces in the range of μN. When imaging within a liquid environment this force is non-existent, thus making imaging in liquids an advantage of contact mode AFM.

It is also possible to operate without the feedback control, in what is known as constant-height contact mode, and is closer to the working principles of a stylus profiler. By holding the tip assembly at a set height, the tip assembly follows a straight path over the surface and the cantilever bends accordingly. This deflection is monitored and used to directly construct a topographical map of the surface. This mode requires additional calibration to carry out effectively, and is decidedly less robust for most applications, but does find use in small, high-speed, high-resolution scans.

The basic data set generated in contact mode is the topography map of the sample, whose data is generated from the movement required by the z-piezoelectric element to maintain a constant cantilever deflection – and thus a constant tip-sample force. This topography image is often published alongside a deflection map, whose data shows the extent of cantilever deflection, specifically

the deflection of the laser dot striking the photodetector as a result of cantilever deformation. The position-sensitive photodetector is split into quadrants A to D, as shown in Fig. 2.1. By comparing the sum of the signal of the top half to the sum of the signal of the bottom half, $(A + B) - (C + D)$, the vertical deflection measurement is calculated.

There is also a third notable signal generated in contact mode AFM caused by the lateral tip-sample forces twisting the cantilever. While rarely used in normal contact mode operation due to the scan direction, this forms the basis of another SPM technique – lateral force microscopy (LFM).

Lateral force microscopy

Often referred to as frictional force microscopy (FFM), LFM is a contact mode variant where the tip is scanned at 90° to the regular AFM scanning direction (essentially swapping the fast-scan and slow-scan directions). In this way, along with vertical deflection of the cantilever due to height variations, the horizontal deflection from torsional twisting of the cantilever is enhanced. This signal is similar to the vertical deflection, but is calculated from the difference between the left and right sides of the photodetector, $(A + C) - (B + D)$. This signal allows the frictional forces on a surface to be measured, allowing a user to identify areas of higher and lower friction as the angle of torsional bending varies during scanning. Figure 2.5 shows the basic principles of LFM operation, and how the LFM signal produced from variations in surface material/friction differ from those produced by variations in topography.

The magnitude of this signal is determined by a variety of factors, such as the frictional coefficient and the topography, but also systematic influences, such as the cantilever's torsional spring constant. LFM is useful for imaging surfaces consisting of inhomogeneous materials, but also allows the user to obtain edge-enhanced images with detail that under normal circumstances would be missed.

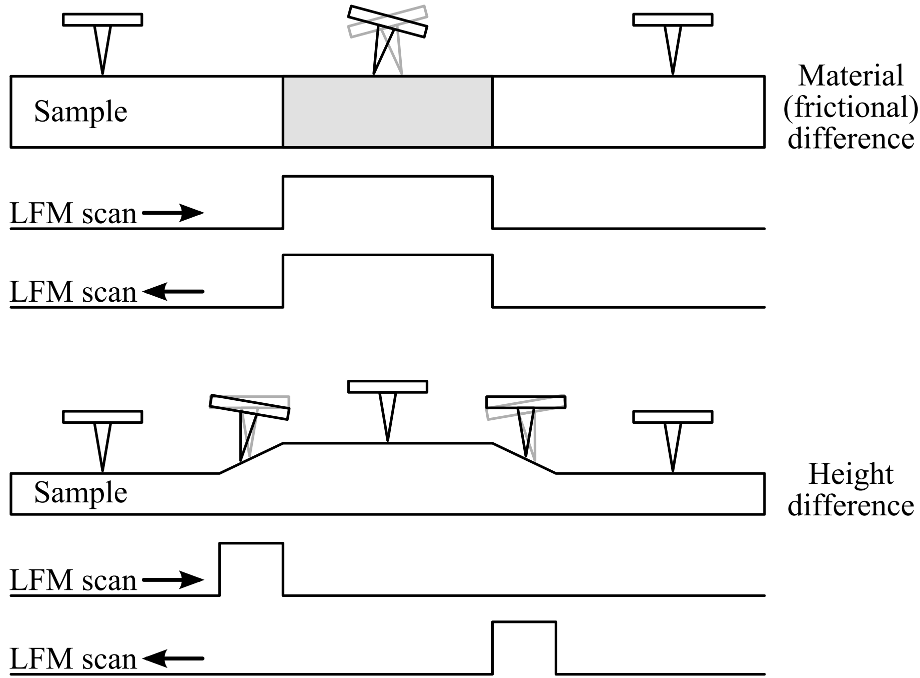


Figure 2.5: Basic operating principle of lateral force microscopy. Cantilever/tip deflections for the scan direction left-right are shown in black, and right-left are shown in light grey.

2.3.2 Dynamic modes

Originally, the vast majority of AFMs made use of contact mode imaging; use of oscillating modes was considered in the original AFM publication [2], but due to the nature of the cantilever used, this mode provided inferior results. However, it wasn't long before these modes were developed further, making use of microfabricated cantilevers, and eventually these so-called dynamic modes became the preferred technique.

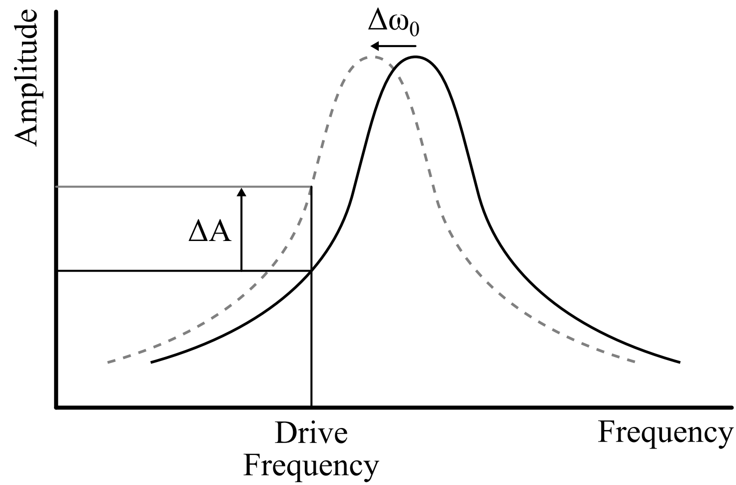
Many SPM techniques are based upon dynamic mode AFM, but all share similar operational principles. A signal is applied to a probe-oscillation piezoelectric element to oscillate the tip at a set amplitude and frequency, usually close to the cantilever's natural resonant frequency, ω_0. The actual movement of the tip depends on the tip-sample interaction, thus the signal recorded by the photodetector will vary depending on the sample area under examination. The oscillation of the cantilever is damped as the tip approaches the sample,

altering its characteristics (amplitude, phase, resonant frequency). This signal is detected by the position-sensitive photodiode, and by comparing the input and output signals it is possible to determine the force acting on the tip. The feedback control system adjusts the z-height accordingly; attractive force gradients acting on the tip reduce the resonant frequency by effectively making the cantilever 'softer', while repulsive gradients have the opposite effect, making the cantilever 'stiffer' and increasing the resonant frequency.

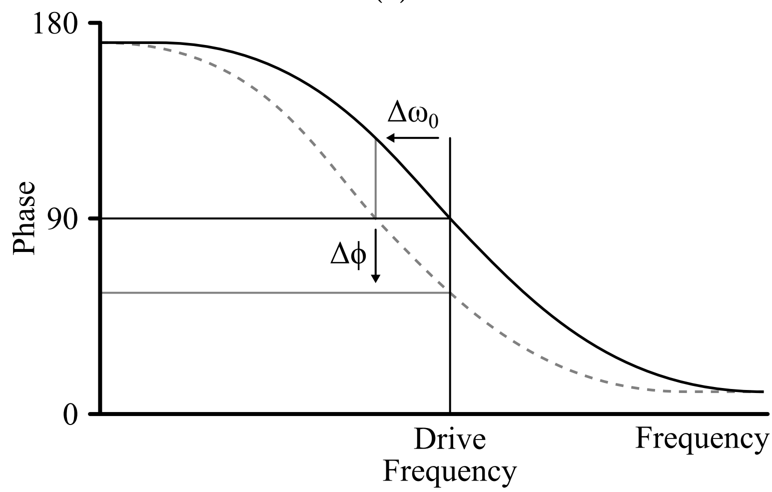
Signal modulation methods

Three main analogue modulation methods are used to compare the input and output signals. Amplitude modulation (AM-AFM) is the most common modulation method used, and works by comparing the change in amplitude between the two oscillating signals for a fixed drive frequency, and adjusting the tip assembly height accordingly. Figure 2.6a outlines how these frequency shifts correspond to changes in oscillation amplitude when in amplitude detection mode. Reducing the resonant frequency of the cantilever increases the oscillation amplitude. By tracking the variations in oscillation amplitude as the tip passes over the sample, an image of the force gradients is produced. This imaging mode is subject to heavier influence of surface topography than other AFM detection modes, as such is it less used in more advanced SPM techniques.

Many newer AFMs also contain the electronics to perform phase modulation (PM-AFM) [6]. While less commonly used due to ambiguous data generation, this method forms the basis of phase imaging, see Fig. 2.6b, a variation of intermittent-contact mode used to examine the material properties of a sample, *e.g.* elasticity. The phase initially crosses the centre line at the drive frequency, 90°. This phase curve decreases with increasing frequency, correctly reflecting the phase lag between the drive voltage and the cantilever response. Force gradients cause a shift in the resonant frequency, giving rise to phase shifts, $\Delta\phi$, which can then be recorded to produce an image of the electric force gradients.



(a)



(b)

Figure 2.6: Schematics of dynamic mode AFM signal modulation methods showing how changes in the resonant frequency of the cantilever can be detected through changes in the amplitude **(a)** and phase **(b)** of the compared input and output signals.

The third signal modulation mode is usually the least commonly used – frequency modulation (FM-AFM). Typically used in ultra-high vacuum (UHV) conditions in non-contact mode, it offers unrivalled sensitivity over other modulation methods, but requires additional equipment to implement. As the resonant frequency of the cantilever is changed by variations in the tip-sample interaction, a frequency demodulator is used to compare the original driving frequency to the new resultant frequency, and adjust the driving frequency to compensate. While it would be simpler to merely record the difference, it is unsafe for the tip to allow such harmonic changes to remain.

Non-contact mode

By applying a low amplitude oscillation to a cantilever, the AFM can be operated in the attractive regime only, often referred to as non-contact, or close-contact mode, see Fig. 2.7. This weak van der Waals attractive force (otherwise known as the London dispersion force or quantum-induced instantaneous polarisation) acting on the tip at a distance of a few nanometres from the sample surface is a result of the correlated movements of electrons in interacting molecules. At these short distances, the proximity of different nearby molecules results in their electrons being repelled from each other, and the molecules become instantaneous dipoles that are attracted to one another.

This mode of operation has the advantage of a low tip-sample force (of the order of pN), and the oscillation phase signal is commonly used to detect cantilever damping, offering higher sensitivity with relatively low amplitudes. This low force interaction allows soft samples to be imaged that would otherwise be damaged by imaging in contact mode. The resonant frequency of the cantilever, ω_0, far from surface is given by:

$$\omega_0 = \sqrt{\frac{k}{m}} \tag{2.9}$$

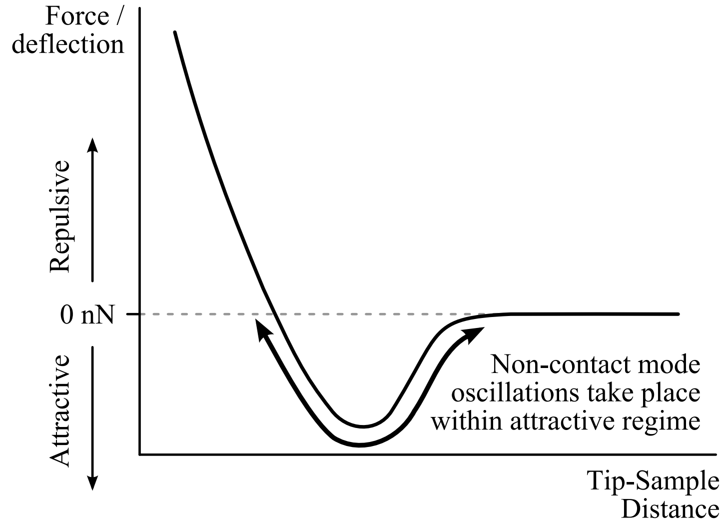


Figure 2.7: Force/deflection diagram showing the attractive force regime acting upon the AFM tip during low amplitude oscillations in non-contact mode.

Where: m is the cantilever mass; and k is the spring constant. To remain oscillating within the attractive regime a stiff cantilever with a high frequency is used, and oscillation amplitudes are kept low at around 10nm. When close to the surface, the tip experiences an additional attractive force [3], F. The new resonant frequency, $\omega\prime_0$, is therefore:

$$\omega\prime_0 = \sqrt{\frac{k - F\prime}{m}} \tag{2.10}$$

Where: $F\prime$ is the derivative of the normal tip-surface force. Non-contact mode AFM can image almost any sample, but is used far less than intermittent-contact mode in ambient conditions due to difficulties arising from the contamination layer, and the larger tip-sample separation which limits the lateral resolution.

The presence of a contaminating fluid layer, usually moisture from the air, can result in non-contact mode imaging the overlaying topography of this layer rather than the sample beneath, masking many important nanoscale features. Furthermore, if the fluid layer is too thick, the tip can become trapped in the adsorbed fluid and can cause unstable feedback. For this reason, non-contact

mode AFM is typically restricted to extremely hydrophobic samples to ensure that any adsorbed fluid layer is at a minimum. However, it is sometimes possible to carefully lower the cantilever until the tip pierces the fluid layer and scan at this height, allowing non-contact mode imaging to be used without any notable interference from the contamination.

Frequency-modulation with applied low amplitude oscillations is used when conducting non-contact mode imaging in ultra-high vacuum (UHV) conditions. This requires additional equipment not present on all AFMs, but provides unrivalled sensitivity, especially on biological samples. To understand why this is a preferred method in UHV conditions, the quality factor, Q, must be considered:

$$\mathrm{Q} = \frac{\omega_0}{\Delta\omega} \tag{2.11}$$

Q is an indicator of the energy loss of the oscillation, and is defined as the ratio of resonant frequency, ω_0, to the full width at half maximum frequency change, $\Delta\omega$. Dynamic mode AFM resolution depends greatly upon the Q factor of an oscillating cantilever. In air this value is approximately 100 for a cantilever with resonant frequency of 300kHz, but at UHV conditions this can be over 500 times greater at 50,000 [7, 8]. The minimum detectable force, δF_{min}, limited by vibrational noise, is [9]:

$$\delta F_{\mathrm{min}} = \sqrt{\frac{2k_{\mathrm{L}}k_{\mathrm{B}}TB}{\omega_0 \mathrm{Q} \langle z_{\mathrm{osc}}^2 \rangle}} \tag{2.12}$$

Where: k_{L} is the normal force constant of the cantilever; k_{B} is the Boltzmann constant; T is the ambient temperature; B is the measurement bandwidth; ω_0 is the resonant frequency of the tip; and $\langle z_{\mathrm{osc}}^2 \rangle$ is the root-mean-square amplitude of the driven cantilever vibration. Comparing this for the two Q values above, an AFM operating in UHV conditions offers a 22 times greater sensitivity than in air. This also explains the greater sensitivity when conducting measurements at cryo-temperatures. However, operating an AFM in AM mode under UHV

conditions is inadvisable as the time scale of amplitude change in AM-AFM, τ_{AM}, is linearly dependent on Q:

$$\tau_{\mathrm{AM}} \approx \frac{2\mathrm{Q}}{\omega_0} \tag{2.13}$$

In FM-AFM, the Q factor and resonant frequency are independent of one another. The time scale of frequency change, τ_{FM}, is instead only inversely proportional to the resonant frequency of the tip:

$$\tau_{\mathrm{FM}} \approx \frac{1}{\omega_0} \tag{2.14}$$

Thus, FM-AFM is a preferred method of modulation when conducting AFM in UHV conditions.

Intermittent-contact mode

Applying a larger amplitude oscillation, up to 100nm, to the cantilever brings the tip closer to the sample than non-contact mode, entering the repulsive regime periodically. This mode is known as intermittent-contact mode, or more commonly 'tapping' mode, and the amplitude signal is most commonly used to detect the damping of the cantilever oscillations. In this mode the tip-sample interaction passes through all force regimes, see Fig. 2.8, from the regime far from the sample (zero-force), through the attractive regime and into the repulsive (contact) regime with each oscillation.

With such a broad range of motion, the tip is brought into contact with the surface intermittently. This introduces the possibility of damaging either the tip or sample, however, by only touching the surface very briefly (and very gently) with each tap, the lateral forces that plague contact mode imaging are virtually eliminated. This proves invaluable for imaging poorly adsorbed samples which could be displaced by imaging in contact mode. The wide range also means that the presence of a contamination layer no longer presents a problem, as the

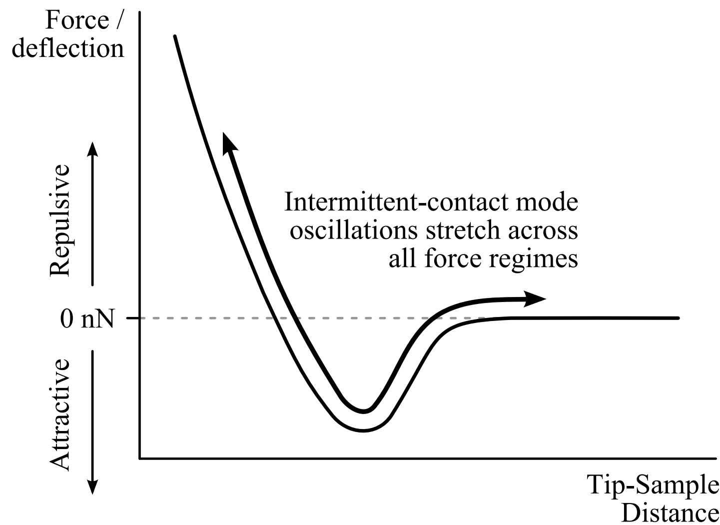


Figure 2.8: Force/deflection diagram showing the range of force regimes acting upon the AFM tip during high amplitude oscillations in intermittent-contact mode.

tip periodically enters and withdraws from this layer with each tap. Due to its versatility, high lateral resolution, and ease of use, tapping mode AFM is commonly used by most AFM practitioners, unless there is a specific reason for using contact or non-contact modes.

In tapping mode AFM, the amplitude signal is mainly used to outline the damping effect from contact with the surface. An amplitude signal map is often shown alongside the topography map, showing where the feedback system has not yet compensated for changes in sample height, similar to the deflection signal map in contact mode AFM. In certain cases it can also be useful to show the phase-shift data, as this can have some uses in distinguishing between materials. This has been further developed into a related technique known as phase imaging AFM. In particular, phase imaging makes use of higher harmonic modes to enhance the imaging capabilities.

2.4 Electric AFM techniques

This section will discuss two relevant electrical SPM modes used throughout this PhD, which are conducted using an AFM.

2.4.1 Electric force microscopy

Electric force microscopy (EFM), otherwise known as electrostatic force microscopy, is a scanning probe technique that examines the local capacitance and potential of a sample [10]. Electrostatic forces are long-range, and as a result EFM has the capability to probe charges trapped slightly below the surface through the shift in surface potential. The sensitivity of EFM is commonly stated to be greater than that of comparable techniques such as scanning capacitance microscopy (SCM), however, it is fundamentally a qualitative technique.

Some advanced AFM modes use multiple-pass/interleave scanning methods, which involve taking multiple scans over the same area to gather different sets of data. One variant of this is the 'lift-mode' technique originally proposed by Bard *et al.* [11], see Fig. 2.9. Lift-mode scanning is most commonly encountered when using electric force microscopy (EFM) or magnetic force microscopy (MFM), allowing the examination of non-planar samples that would otherwise be immeasurable due to their shape.

The first scan in lift-mode is conducted to obtain the topography, usually conducted in intermittent-contact mode. The tip is raised by a set distance from the surface to bring the electrostatic forces into dominance. The van der Waals attractive forces are proportional to $1/r^6$, but electrostatic forces are proportional to $1/r^2$, thus moving the tip away from the sample surface will decrease the van der Waals forces and leave the electrostatic forces dominant.

A second, non-contact mode scan follows the recorded path of the topography from the first scan at this raised height. In theory, this should be a path of constant van der Waals force, without any topographical interactions. The only

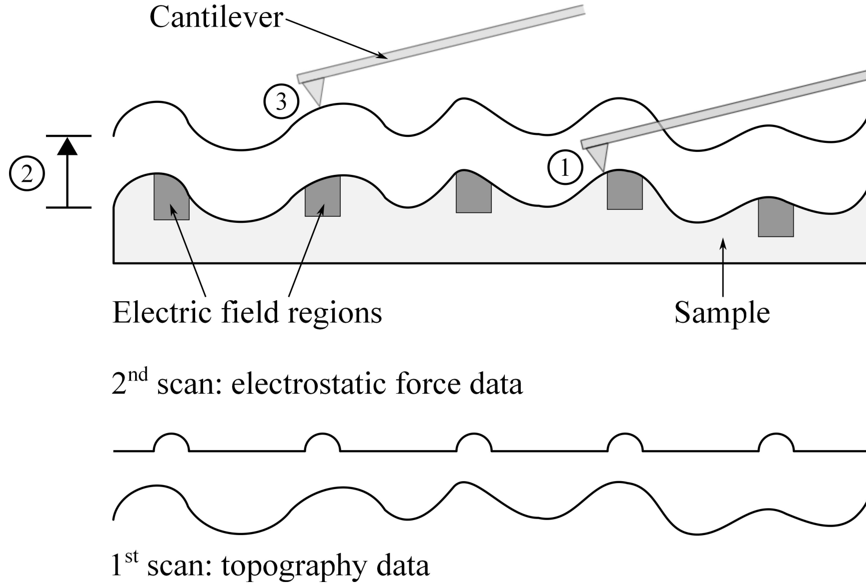


Figure 2.9: Schematic of how lift mode EFM technique takes place. Initially the topography is scanned (1). The cantilever is then lifted by a set amount (2). A second EFM scan at this constant height accounts for topographical features (3).

tip-sample interactions should result from changes in the electrostatic force. The tip is scanned without feedback, and can be biased to enhance the resultant images. Modelling the tip-sample interaction as a capacitor with capacitance, C, with a tip-sample potential difference, ΔV, and tip-sample separation, z, the total electrostatic potential energy stored in this capacitor, U_E, is:

$$U_E = \frac{1}{2} C \Delta V^2 \tag{2.15}$$

Where: C is the capacitance; and V is the electric potential difference. Taking the derivative of U_E with respect to the tip-sample separation distance, z, gives us an expression for the electrostatic force interaction:

$$F_E(z) = -\frac{\partial U_E}{\partial z} \tag{2.16}$$

$$= -\frac{1}{2} \Delta V^2 \frac{\partial C(z)}{\partial z} \tag{2.17}$$

For an oscillating cantilever travelling over a charged region, the resonant frequency will be altered as it is subjected to an additional force due to the electric field gradient from the charged region beneath, see Eq. (2.10). An attractive electrostatic force gradient acting upon the tip will result in a reduction in the resonant frequency of the cantilever, and a repulsive gradient will result in an increase in the resonant frequency.

This change in resonant frequency is proportional to changes in tip-sample capacitance as a function of the second derivative of tip-sample separation. As long as there is a non-zero potential between the tip and surface, the frequency, and thus the amplitude and phase of the oscillation, are sensitive to the capacitance of the surface. EFM can be conducted using all three modulation modes: amplitude, phase, and frequency. Due to a faster response time, phase or frequency modulation are preferred over amplitude modulation when conducting EFM.

The best candidates for EFM examination are samples with smooth topography and large contrasts in the electric force gradient due to material differences, trapped charge, or regions held at substantially different potentials (>1V). Applying a bias to the tip or sample will enhance the electric field response and increase the contrast due to material differences in the EFM scan, however, large biases should be avoided when imaging samples with permanent electric fields, *e.g.* those with trapped charge.

Due to the nature of the tip-sample interaction, when imaging samples with a rough surface topography, EFM may be undesirable. Sharp surface features will concentrate local electric force gradients, and appear in the EFM image as edge artefacts similar to the topography image. These artefacts can persist, even using lift-mode; furthermore, since the tip-surface interaction is taken to be vertical, nearby high features that induce near-vertical interactions with the tip will also have a detrimental effect.

2.4.2 Scanning Kelvin probe microscopy

Scanning Kelvin Probe Microscopy (SKPM), otherwise known as Kelvin Probe Force Microscopy (KPFM) or Scanning Surface Potential Microscopy (SSPM), was originally introduced by Nonnenmacher *et al.* [12] in 1991 and has since found many diverse applications for the direct and quantitative measurement of surface potential distributions across a wide range of fields. An SKPM can typically measure the work function of materials down to a few hundred meV [13], has a sensitivity of a few millivolts and a lateral spatial resolution of an order of nanometres – the typical spatial resolution of electrical AFM techniques.

A scanning Kelvin probe measures the work function, Φ, of a surface. This is defined as the energy required to move an electron from the Fermi level, E_{F}, to a point immediately outside the surface – the vacuum level, E_{vac}. The Fermi level is the energy level at which the probability of finding an occupied state is 0.5. In conducting materials the Fermi level is located within the conduction band, but for insulating materials and intrinsic semiconductors it is located inside the band gap. If two materials have different Fermi levels, making an external contact between them allows electrons to transfer across, equalising the two Fermi levels; this transfer creates what is known as a contact potential difference (CPD) between the two materials. The CPD between two materials, *e.g.* between the AFM tip and the sample surface, is defined as:

$$V_{\mathrm{CPD}} = \frac{\Delta\Phi}{e} \tag{2.18}$$

$$= \frac{\Phi_{\mathrm{tip}} - \Phi_{\mathrm{sample}}}{e} \tag{2.19}$$

Where: Φ_{tip} and Φ_{sample} are the work function values of the tip and sample, respectively; and e is the elementary charge. Thus, if an AFM tip and a semiconductor sample with different work functions are held in close proximity, an

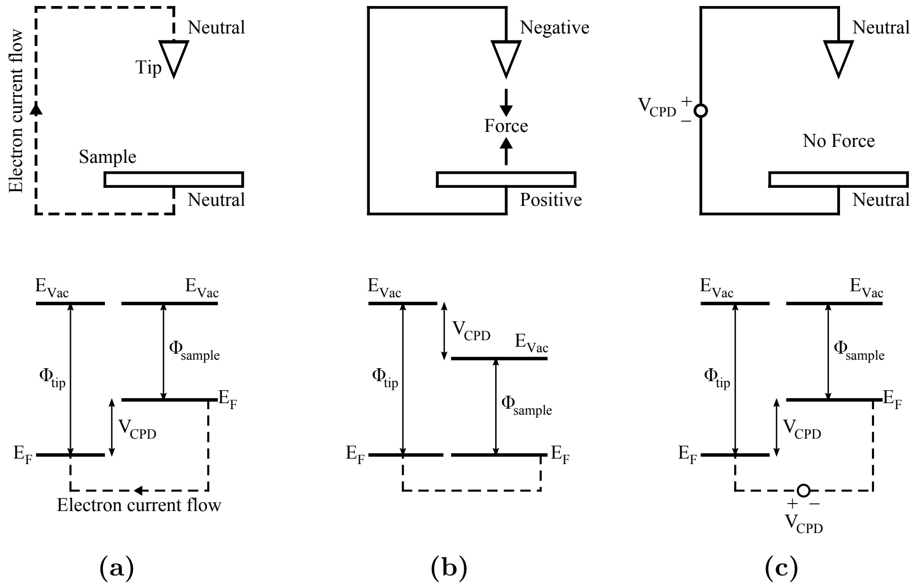


Figure 2.10: Definition and basic measurement setup of contact potential difference. Connection occurs **(a)**, an electrostatic force develops **(b)**, and is nullified by an external bias **(c)**

electrostatic force will develop between them due to the contact potential difference, V_{CPD} – this is outlined in the energy level diagrams and basic schematics shown in Fig. 2.10.

When the tip and sample are separated by a large distance, they are both neutral, *i.e.* their local vacuum levels are aligned but their Fermi levels differ. As they are brought into close proximity, electrical connection occurs and the Fermi levels will align due to electron flow, see Fig. 2.10a. The tip and sample become charged, and their local vacuum levels are now different. Due to the opposing charges on the tip and sample in response to electrical connection, an electrostatic force develops, shown in Fig. 2.10b. This electrostatic force can be nullified by applying an external bias between the tip and sample, shown in Fig. 2.10c. The magnitude of this applied bias is the CPD. The sign of the applied voltage, V_{CPD}, will be negative if the nullifying voltage is applied to the tip, and positive if applied to the sample.

Treating the tip-sample system as a parallel plate capacitor, as in EFM above, the total electrostatic potential energy stored in the capacitor, U_E, is:

$$U_E = \frac{1}{2} C \Delta V^2 \tag{2.20}$$

Where: C is the capacitance; and V is the electric potential difference. As above, taking the derivative of U_E with respect to the tip-sample separation distance, z, gives us an expression for the force:

$$F_E(z) = -\frac{\partial U_E}{\partial z} \tag{2.21}$$
$$= -\frac{1}{2} \Delta V^2 \frac{\partial C(z)}{\partial z} \tag{2.22}$$

When the tip has both DC and AC voltages applied:

$$V_{\text{tip}} = V_{DC} + V_{AC} \sin(\omega t) \tag{2.23}$$

The voltage difference between tip and sample is thus:

$$\Delta V = V_{\text{tip}} \pm V_{\text{CPD}} \tag{2.24}$$
$$= (V_{DC} \pm V_{\text{CPD}}) + V_{AC} \sin(\omega t) \tag{2.25}$$

As previously mentioned, the $\pm$ sign for the contact potential difference relates to whether the voltage is actually applied to the tip ($-$) or sample ($+$). Substituting this into the expression for the z-component of the electrostatic force:

$$F_E(z,t) = -\frac{1}{2} \frac{\partial C(z)}{\partial z} \left[(V_{DC} \pm V_{\text{CPD}}) + V_{AC} \sin(\omega t) \right]^2 \tag{2.26}$$

Expanding the quadratic expression for the voltage difference gives three components:

$$F_E(z,t) = F_{DC} + F_\omega + F_{2\omega} \tag{2.27}$$

$$F_{DC} = -\frac{1}{2}\frac{\partial C(z)}{\partial z}(V_{DC} \pm V_{\text{CPD}})^2 \tag{2.28}$$

$$F_\omega = -\frac{\partial C(z)}{\partial z}(V_{DC} \pm V_{\text{CPD}})V_{AC}\sin(\omega t) \tag{2.29}$$

$$F_{2\omega} = \frac{1}{4}\frac{\partial C(z)}{\partial z}V_{AC}^2(cos(2\omega t) - 1) \tag{2.30}$$

The DC component of the electric force, F_{DC}, contributes to the topographical signal; F_ω is used to measure the CPD in SKPM; and the final contribution, $F_{2\omega}$, can be used when conducting scanning capacitance microscopy (SCM).

Many authors use the F_ω contribution shown above, however, it is only strictly correct for a metallic tip-sample system. When examining semiconductor samples another model is needed to explain the tip-sample interaction. Hudlet *et al.* presented a straightforward one-dimensional analysis of the electrostatic force interaction between tip and surface in the case of a semiconductor sample [14, 15] and have shown the force at a frequency, ω, can be expressed as:

$$F_\omega = \frac{Q_{\text{s}}}{\epsilon_0}C_{\text{eff}}V_{AC}\sin(\omega t) \tag{2.31}$$

Where: Q_{s} is the semiconductor surface charge; C_{eff} is the effective capacitance of the cantilever tip-air/vacuum-semiconductor surface system; and ϵ_0 is the dielectric constant. For a metallic sample:

$$Q_{\text{s}} = -C\,(V_{DC} \pm V_{\text{CPD}}) \tag{2.32}$$

Where: C_{eff} is replaced by C, the capacitance between tip and metallic sample. In such a scenario, the two variations of F_ω equate and the model is still valid.

The input of the feedback loop depends on the detection mode used. For amplitude modulation, given that the cantilever's response at its resonant frequency is significantly stronger than at the second harmonic, the cantilever's vibrational amplitude, A, is to a first approximation directly proportional to the amplitude of F_ω.

$$A \propto \frac{\partial C(z)}{\partial z}(V_{DC} \pm V_{\mathrm{CPD}})V_{AC} \tag{2.33}$$

For frequency modulation, a force gradient is the input for the frequency demodulator feedback loop, Δf:

$$\begin{aligned} \Delta f &\propto \frac{\partial F_\omega}{\partial z} && (2.34) \\ &\propto -\frac{\partial^2 C(z)}{\partial z^2}\left(V_{DC} \pm V_{\mathrm{CPD}}\right) V_{AC} \sin(\omega t) && (2.35) \end{aligned}$$

Both modulation modes can be used to detect a sample's topography and the CPD simultaneously, although most current 'turnkey' commercial systems tend to use lift-mode to minimise the effects of topography on the potential scan image. The first scan measures the sample topography in tapping mode – the cantilever is vibrated near to its resonant frequency by a small piezoelectric element. The cantilever is then raised a specified distance and the piezoelectric element that normally mechanically vibrates the cantilever (as in EFM and other lift-mode techniques) is turned off. Instead, an AC voltage is superimposed on a modulated DC tip voltage, resulting in an oscillating electrostatic force, which in turn induces an oscillation of the cantilever at frequency ω.

The force on the cantilever depends on the product of the AC drive voltage and the DC voltage difference between tip and sample. When the tip and sample are at the same DC voltage the cantilever will feel no oscillating force. The effective local surface potential (CPD) of the sample is determined by adjusting

the DC voltage on the tip, using a feedback loop, until the oscillation amplitude becomes zero. When the amplitude is at zero, the tip voltage will be equal to the unknown surface potential, thus the voltage applied to the cantilever tip can be recorded and plotted to give an image of the surface potential. The operation of this feedback loop is what truly differentiates SKPM from EFM – the lack of feedback used during the lift-mode EFM scan is also the reason why EFM is sometimes referred to as 'open-loop' SKPM.

It should be noted that an AC drive voltage above 2V is necessary to obtain sufficient sensitivity for most materials. It should also be noted that when dealing with semiconductor samples, voltage-induced band bending is also a possibility, thus AC voltages as low as possible should be used.

By nullifying the electrostatic force to find the local CPD of a sample, SKPM is typically more successful at imaging samples with sharper surface topography than EFM. Usually EFM can be carried out on a standard AFM, but additional equipment is required to equilibrate the tip-surface potential when conducting SKPM.

Bibliography

[1] R. Young, J. Ward, and F. Scire, "The topografiner: An instrument for measuring surface microtopography," *Review of Scientific Instruments*, vol. 43, no. 7, pp. 999–1011, February 1972.

[2] G. Binnig, C. F. Quate, and C. Gerber, "Atomic Force Microscope," *Physical Review Letters*, vol. 56, no. 9, pp. 930–933, Mar. 1986.

[3] Y. Martin, C. Williams, and H. Wickramasinghe, "Atomic force microscope - force mapping and profiling on a sub 100-åscale," *Journal of Applied Physics*, vol. 61, no. 10, pp. 4723–4729, November 1987.

[4] G. Meyer and N. Amer, "Erratum: Novel optical approach to atomic force microscopy," *Applied Physics Letters*, vol. 53, no. 24, pp. 2400–2402, 1988.

[5] O. Sahin, S. Magonov, C. Su, C. Quate, and O. Solgaard, "An atomic force microscope tip designed to measure time-varying nanomechanical forces," *Nature Nanotechnology*, vol. 2, no. 8, pp. 507–514, 2007.

[6] T. Fukuma, J. Kilpatrick, and S. Jarvis, "Phase modulation atomic force microscope with true atomic resolution," *Review of Scientific Instruments*, vol. 77, no. 12, pp. 123 703–123 708, 2006.

[7] F. Giessibl, "Advances in atomic force microscopy," *Reviews of Modern Physics*, vol. 73, no. 3, pp. 949–983, Jul 2003.

[8] T. Albrecht, P. Grütter, D. Horne, and D. Rugar, "Frequency modulation detection using high-q cantilevers for enhanced force microscope sensitivity," *Journal of Applied Physics*, vol. 69, no. 2, pp. 668–674, Sep 1991.

[9] W. Melitz, J. Shen, A. Kummel, and S. Lee, "Kelvin probe microscopy and its application," *Surface Science Reports*, vol. 66, no. 1, pp. 1–27, Jan 2011.

[10] B. Rodriguez, A. Gruverman, and R. Nemanich, "Nanoscale characterization of electronic and electrical properties of iii-nitrides by scanning probe microscopy," in *Scanning Probe Microscopy - Electrical and Electromechanical Phenomena at the Nanoscale*, S. Kalinin and A. Gruverman, Eds. Springer, 2007, vol. 3.

[11] C. Lin, F.-R. Fan, and A. Bard, "High resolution photoelectrochemical etching of n-gaas with the scanning electrochemical and tunneling microscope," *Journal of The Electrochemical Society*, vol. 134, no. 4, pp. 1038–1039, 1987.

[12] M. Nonnenmacher, M. P. O'Boyle, and H. K. Wickramasinghe, "Kelvin Probe Force Microscopy," *Applied Physics Letters*, vol. 58, no. 25, pp. 2921–2923, 1991.

[13] S. Sadewasser, "Surface potential of chalcopyrite films measured by KPFM," *Physica Status Solidi (a)*, vol. 203, no. 11, pp. 2571–2580, Sep. 2006.

[14] S. Hudlet, M. Jean, B. Roulet, J. Berger, and C. Guthmann, "Electrostatic forces between a metallic tip and semiconductor surfaces," *Microscopy Microanalysis Microstructures*, vol. 5, no. 4-6, pp. 467–476, 1994.

[15] S. Hudlet, M. S. Jean, B. Roulet, J. Berger, and C. Guthmann, "Electrostatic forces between metallic tip and semiconductor surfaces," *Journal of Applied Physics*, vol. 77, no. 7, pp. 3308–3314, 1995.

Chapter 3

SIM cards

With the AFM and its subset of electrical analysis techniques covered in the previous chapter, this chapter covers the background information, exhibit processing steps, and AFM analysis of SIM card EEPROM memory arrays with the initial aim of developing a viable method of data recovery for damaged electronic memory devices.

Before processing and analysis of an exhibit can be undertaken, it is useful to understand various features of smart cards. The topics discussed in this chapter include: the physical structure of a contact smart card; an overview of the microprocessor structure; a deeper explanation of the operational mechanisms of the EEPROM memory arrays; EEPROM/flash memory data retention characteristics; and finally, an overview of data held within a SIM card that is important to forensic examiners.

3.1 History of smart cards

It is worth noting certain points on the nomenclature when researching the history of smart cards. Smart cards, as we know them today, came under various names while first being developed and patented: 'chip cards', 'automated chip cards', 'memory cards', 'self-programmable one-chip microcomputers (SPOM)',

'microprocessor smart cards', etc. These names gradually died out and today we have settled on 'smart cards'.

The automated chip card (carrying a microchip within a plastic card) was conceived and filed for patent in 1968 by Helmut Gröttrup and Jürgen Dethloff and granted in 1982. Smart cards were first patented in 1970 by Kunitaka Arimura[1]. Four years later the patent for the IC card (later renamed the 'smart card') was filed by Roland Muréno, who specifically patented several functional aspects of the smart card between 1974 and 1976, focusing mainly on the memory access technology. Michel Ugon invented the first microprocessor smart card in 1977, and applied for the US patent in 1978 [1], he held numerous patents (>1000) in this area with Honeywell Bull.

The first mass use of smart card technology was a credit card payment system in France to combat fraudulent use in restaurants, closely followed by France Telecom introducing telephone payment cards in pay phones in 1983. The large-scale integration of microchips into all French debit cards was completed in 1992, and in 1993 the current standard system for credit and debit smart card payments began, and is known as EMV after the initial letters of the three founding companies: Europay, MasterCard, and VISA. EMV developed a set of further specifications covering the core functions of a banking smart card in far greater detail than ISO 7816 which defines the standards required of a general integrated circuit smart card.

Smart cards typically come in two varieties: 'memory cards' and 'microprocessor cards'. The former is dedicated to data storage in a non-volatile fashion and contains limited security logic. These are ubiquitous today in the form of Secure Digital (SD) memory cards (the successor to the MultiMediaCard (MMC) memory cards). Microprocessor cards, however, contain both volatile and non-volatile memory and can compute cryptographic functions for security authentication. By far the most commonly found microprocessor cards found

[1] www.cardwerk.com/smartcards/smartcard_history.aspx

today are in the form of mobile phone subscriber identity module (SIM) cards, bank cards, contactless public transport cards, and building access cards.

The first commercial SIM cards were manufactured by Giesecke & Devrient for the Finnish mobile network, Radiolinja in 1991. The introduction of this smart card-based SIM for use in the GSM mobile phone system in the 1990s created a major boom in smart card use throughout Europe. ETSI, the European Telecommunications Standards Institute, is responsible for standards covering smart cards used in public and cellular telecommunications systems. Put simply, the SIM card is a type of smart card containing a service-subscriber key used to identify a subscriber using a mobile phone handset to the network. With their cryptographic functions, these SIM smart cards are very similar to other smart cards commonly available, such as building security access cards and even 'Chip-and-PIN' credit/debit cards.

Because of its widespread use, the term 'SIM Card' has become synonymous with any variant of subscriber smart card used in mobile phones. Nevertheless it should be clarified that 'SIM' was originally both the physical smart card and the software application for subscriber verification on a 2G GSM network. With newer Universal Mobile Telecommunications System (UMTS) networks there was need for an upgraded SIM application. The physical SIM card thus became known as a Universal Integrated Circuit Card (UICC), while the original subscriber verification software for 2G GSM networks continued to be called SIM. The software application for use on 3G UMTS networks is called Universal Subscriber Identity Module (USIM) and is installed alongside the SIM software allowing backward compatibility on older 2G GSM-only networks. A common mistake with many mobile phone networks is to mislabel the newer UICC-type smart cards as 'USIMs' when USIM really only applies to one of the applications on the UICC smart card.

According to the GSM Association, by mid-2009 approximately 80% of the global mobile market was held by GSM technology. With over 6 billion global

Image removed for
copyright reasons

Figure 3.1: A graph showing the global mobile subscription rates, and divisions between developed and developing world countries. From: *'Measuring the Information Society – 2012'*, U.N. International Telecommunication Union.

mobile phone connections as of 2011, it has become commonplace in most developed and developing countries to have a mobile phone. Many countries now have biometric passport technology in place. This uses contactless smart cards built into the back page of the passport or inside an identity card containing information about the bearer. In Malaysia, the compulsory national identity scheme 'MyKad' has been rolled out successfully and contains 8 different applications: identity card; driving licence; passport; health information; ATM integration; road toll and public transport payment system; e-cash; and a digital security certificate on more modern versions.

A report by the UN International Telecommunication Union (ITU) entitled *'Measuring the Information Society – 2012'* recently found that there are almost as many mobile phone subscriptions in the world as people. A graph of worldwide subscription rates per 100 inhabitants is shown in Fig. 3.1. Between 2010 and 2011 alone there was an increase of 600 million mobile subscriptions, mainly in developing countries.

3.2 SIM card structure

Smart cards come in various shapes and sizes, see Table 3.1. These are outlined in ISO 7810 standard, and for Micro/Nano-SIM cards, ETSI TS 102 221 v9.0.0.

A cross section of the chip module area of a SIM card with major components labelled is shown in Fig. 3.2. The plastic card body is usually polyvinyl chloride (PVC) but can also be made from other thermoplastics such as acrylonitrile butadiene styrene (ABS). The chip module area is a milled or moulded cavity inside the main card body into which the chip and contact pads are mounted during production.

Table 3.1: Common smart card standard dimensions and examples of their respective uses.

Name	Length (mm)	Width (mm)	Thickness (mm)	Usage
SIM, ID-1[a]	85.60	53.98	0.76	Full-sized SIM/smart cards, credit/debit cards, ID cards
ID-2[a]	105	74	0.76	German ID cards issued prior to Nov 2010
ID-3[a]	125	88	0.76	Passports, visas
Mini-SIM, ID-000[a]	25	15	0.76	Mini-SIM cards
Micro-SIM, Mini-UICC[b]	15	12	0.76	Micro-SIM cards, in use from 2003 onwards
Nano-SIM[c]	12.30	8.80	0.67	Newly-introduced Nano-SIM cards, 2012

a. ISO/IEC 7810:2003
b. ETSI TS 102 221 V9.0.0
c. ETSI TS 102 221 V11.0.0

Prior to installation into the cavity, contact wires are bonded between the chip and the external contact pads. The chip module is then encapsulated within an epoxy moulding compound (EMC), traditionally called a ‘glob top’. Inorganic fillers such as fused silica are commonly used to lower the thermal expansion of EMC and various types of rubber, known in the industry as ’stress modifiers‘, are incorporated into the EMC to reduce the brittle nature of the epoxy and

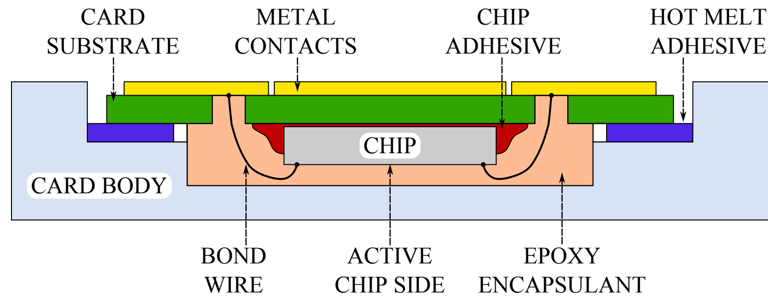


Figure 3.2: Enlarged cross section of the chip module area of a SIM card.

improve toughness. EMCs contain a variety of compounds in ratios similar to those outlined in Table 3.2.

The encapsulant surrounding the chip is not always made from epoxy; it can also be ceramic-based or plastic-based. The exact recipe for the encapsulant can be adjusted for a wide range of physical properties to suit different purposes, *e.g.* thermally conductive fillers may be added to improve the thermal conductivity of the set polymer/ceramic composite [2].

One of the challenges facing any forensic investigator looking to examine a SIM card die is that they must first remove any encapsulating material with minimal harm to the die and without altering the stored data held within. Some of the more commonly used methods in failure analysis, such as plasma etching, are therefore unsuitable for forensic investigation.

Table 3.2: Typical constituents of epoxy moulding compounds used in the semiconductor industry [3].

Material	Percentage (by weight)
SiO_2 / Al_2O_3 / AlN Filler	60–80
Epoxy Resin	5–20
Phenol-based Hardener	5–10
Brominated Epoxy Resin	1–5
Tertiary Amine-Phosphorous	<1
Epoxy Silane	<1
Stearic Acid	<1
Wax	<1
Carbon Black	<1
Silicone / Synthetic Rubber	<1

3.2.1 SIM card contact pads

As can be seen in Fig. 3.3, there are 8 contact pads on a smart card. In a mobile phone SIM card, pads C4 & C8 are not used and are thus not always present; these are reserved for Data +/− in USB compatible smart cards.

C1 (VCC)

Supply voltage, VCC.

GSM/UMTS SIM cards (TS 102 221)

Class A -	5V±10%	1–5MHz	10mA @ 5MHz (operating state)
Class B -	3V±10%	1–5MHz	7.5mA @ 5MHz (operating state)
Class C -	1.8V±10%	1–5MHz	5mA @ 5MHz (operating state)

Payment Cards (EMV)

	5V±10%	1–5MHz	50mA

C2 (RST)

Reset input used to switch the smart card microcontroller on/off.

C3 (CLK)

Clock input delivers an external clock signal (1–10MHz) that is used as a system clock for both the microcontroller and serial communication.

C4 (AUX1)

Auxiliary contact; USB devices: Data +

C5 (GND)

Ground.

C6 (VPP)

Formerly the EPROM programming voltage, VPP (around 21V was required [4]). This is not typically used any more since modern cards generate the programming voltage on-chip using a charge pump fed by VCC.[2]

[2]Some modern smart cards use C6 as a communication port for Near Field Communication via Single Wire Protocol.

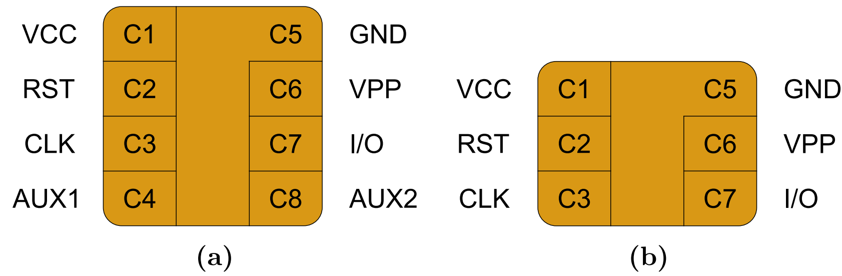


Figure 3.3: Smart card electrical contact pad arrangement for 8 pads **(a)**, and 6 pads **(b)**.

Image removed for
copyright reasons

Figure 3.4: Smart card electrical contact pad designs specific to manufacturers. Ca. 2000. From [4].

C7 (I/O)

Input/Output for serial communication[3].

C8 (AUX2)

Auxiliary contact; USB devices: Data –

SIM card contacts can come in a wide variety of patterns, often used to denote different manufacturers and generations of card/chip. However, the contact pads always have a 3×2 or 4×2 arrangement. If the 2×4 pad layout is used, it is shifted to maintain the 6 existing contact point locations in card readers. ISO 7816 Part 2 dictates only the location for which electrical contact between card and terminal take place, allowing for manufacturers to differentiate their products using unique patterns or artwork. Figure 3.4 shows some examples of module contact pad design patterns.

[3]The serial I/O communication protocols are outlined in Standard ISO 7816, Part 4, Appendices A & B.

3.3 SIM card microcontroller components

Most smart card microcontroller chips are built around the von Neumann architecture, shown in Fig. 3.5, whereby data within the chip is passed through a central bus under control of the security logic. This security logic controls access to different areas of memory within the device. Coupled with the memory partitioning by function, and this forms the basis of on-chip smart card security.

The embedded EEPROM memory array can be sub-divided into two or more areas. One area is immediately accessible upon power cycling and often forms part of the response to the smart card terminal, this is known as Answer-To-Reset (ATR). The rest of the memory array partitions are only accessible once the card has received one of the valid security codes, be it a manufacturer code, a code from a program reading data from the card, or a user PIN code (Personal Identification Number).

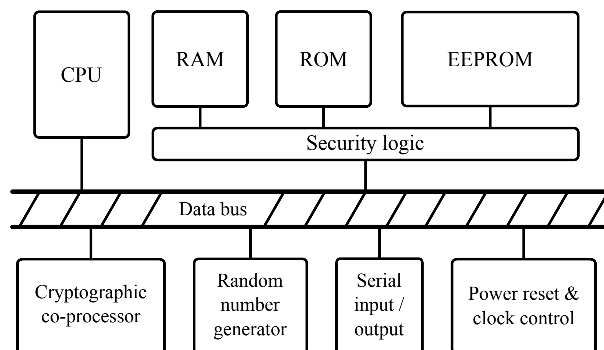


Figure 3.5: A simplified schematic showing von Neumann smart card microcontroller architecture.

The SIM card microcontroller contains all the components necessary for processing the encryption algorithms needed for authenticating network access, and also for modifying and accessing the onboard data storage array of user data. More advanced smart card chips can contain a second processor dedicated solely to cryptography. An example of such a chip, the Infineon SLE 66CX160S, is shown in Fig. 3.6 with main onboard components labelled.

The SLE66 family of chips dates from around 2000 and is evaluated under

Image removed for
copyright reasons

Figure 3.6: Components labelled on an Infineon SLE 66CX160S smart card microcontroller. After: [5].

Common Criteria to EAL5+. This chip is still one of the core chips used in GSM and EMV fields, as well as other applications such as Trusted Platform Module (TPM) as used in Xbox360 games consoles to counter a user modifying the circuitry to allow software from multiple regions, as well as pirated versions, to be used (known as 'chipping' the console).

The 'Common Criteria for Information Technology Security Evaluation' (often shortened to 'Common Criteria') is an international standard for computer security certification. This includes rigorous independent assessment of a product's claims on security capabilities. The 'Evaluation Assurance Level' is a 7 level system denoting the level of testing conducted to guarantee the security claims and functionality made about a particular product[4].

Cryptographic smart cards such as these will typically incorporate a wide range of security countermeasures such as: memory address scrambling; detection circuits to block access to the card after multiple failed PIN attempts, power-circuit manipulation, or clock manipulation; or by incorporating conductive plates across different layers on the surface of the die that will wipe stored data if charged by imaging in an electron microscope.

[4]A common misconception is that a higher EAL denotes a more secure product when in fact it means it has been more rigorously tested to abide by security claims.

3.4 Smart card memory

Memories come in various forms, but essentially can be split into two main groups: volatile and non-volatile. Volatile memories, such as RAM in a PC rely on power to store data. Volatile memories are commonly the fastest type of memory for read/write applications, but they do require power to hold data – upon disconnecting the power source the memory is cleared. Non-volatile memories do not rely on an external power supply to hold the data; even after power loss the data charges are secured in isolated regions within the memory array.

As previously mentioned, a cryptographic smart card contains a variety of memory types: working memory in the form of RAM, pre-programmed smart card applications stored on a ROM array, and user data stored in an EEPROM array or its flash variant. EEPROMs are commonly used to store the majority of the data on a smart card. EEPROM is a type of non-volatile memory first developed at Intel in 1978 by George Perlegos; it was an improvement over earlier designs, incorporating a thin gate oxide layer to enable on-device electrical erasing, rather than cumbersome UV-erasing required on EPROM devices.

EEPROM, unlike RAM, can retain data after the power is removed from the device, making them ideal for many applications. Flash memory, a close relative to the EEPROM, is the most popular modern non-volatile memory available today. The difference between EEPROM and flash lies in the cell design; while EEPROM has the ability to erase individual bits, flash memory only erases data one 'page' at a time. This greatly improves the speed of data writing, a notable drawback with EEPROM arrays. EEPROM features a 2T cell, the pair of transistors consisting of one access transistor and one floating gate transistor. Flash has a 1T cell size, allowing larger, more densely packed arrays.

3.4.1 Floating gate transistors

Data is held within an EEPROM array in a series of floating gate transistors (FGT). Figure 3.7, a schematic of a FGT, shows the main difference between this type of stacked gate transistor and a conventional metal-oxide-semiconductor field-effect transistor (MOSFET) – the inclusion of a second gate beneath the control gate. This gate, termed a 'floating' gate, is electrically isolated from surrounding structures by an insulating inter-polysilicon dielectric (IPD) layer above, and a thinner tunnel oxide layer below. The first floating gate device was proposed by Kahng and Sze in 1967 [6].

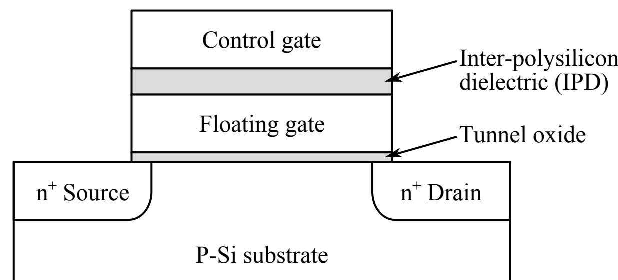


Figure 3.7: Simplified diagram of a stacked floating gate transistor.

Doping specified wells in the silicon to be n-type or p-type gives a source and drain for the floating gate transistor. A tunnel oxide layer of 5–12nm of SiO_2 is deposited onto the silicon to form the thin tunnel oxide layer. Doped polycrystalline silicon (polysilicon) is used to form the floating gate, with a separation layer of SiO_2 or $SiO_2/Si_3N_4/SiO_2$ (ONO) composite layer forming the second insulative IPD layer. Charge stored within the floating gate acts to screen the effects of the control gate's electric field, varying the threshold voltage (V_T) of the transistor and allowing high or low values to be represented logically. In practical devices, the bottom tunnel oxide is of ≈8nm, while the IPD typically has a thickness of ≈14nm. Aluminium is typically used in higher layers for longer interconnects.

Originally EEPROM cells consisted of two transistors; each FGT was partnered with a regular MOSFET known as an access transistor. These were used

to individually select their corresponding FGTs for programming/erasing operations. In the mid-1980s a new variant of the EEPROM was produced – flash memory. For flash devices, erasing is only carried out over large areas of the memory array, not on a single-cell basis as in earlier EEPROM devices. Without the need to erase individual FGTs, more complex architecture could be developed and the cell size could be reduced, allowing flash cells to be 2–3 times smaller than EEPROM cells, and correspondingly more densely packed into an array.

3.4.2 Operation mechanisms

As mentioned previously, the floating gate in the FGT stores charge and is electrically isolated by dielectric material on all sides. Two main methods are used by EEPROM devices to inject charge into these floating gates – hot carrier injection (HCI) or Fowler-Nordheim[5] tunnelling (FNT). The amount of charge injected into floating gates varies between devices, but generally between 10^3–10^5 electrons. Hot carrier injection takes place at the drain-side and is thus a non-uniform method; however, FNT can operate either at the drain-side (non-uniform) or across the entire channel (uniform). This second, uniform version of FNT results in lower oxide stresses, and is the most common method of removing charge from the floating gate.

Uniform write and uniform erase technology was proposed by Kirisawa *et al.* [7]. Whichever operation is used for programming, FNT is used in erasing the floating gate. An excellent explanation of the working of a FGT is given by Sze and Ng in their textbook *'Physics of Semiconductor Devices'* [8].

The high voltages needed during programming and erasing were previously supplied externally (see contact pad C6 in Fig. 3.3). They had a 5V supply for read operations (VCC) and 12V supply for write operations (VPP). By 2007, nearly all EEPROM and flash memory ICs contained integrated charge pumps

[5]Named after Ralph H. Fowler and Lothar Wolfgang Nordheim

to internally generate the high voltages required, and only a single VCC supply voltage, typically 1.8V or 3.3V, is required.

It is worth taking note at this point of the nomenclature relating to the processes of moving charge in/out of floating gate transistors in EPROM (and later) memories. It is all too common to find mistakes which, while minor, cause some confusion and many a publication, including some textbooks, have minor errors in this regard. The root of this appears to stem from a misuse of the terms 'programme', 'write', 'erase', and their resultant logical states 1 and 0.

The correct definitions are as follows: injecting charge into the floating gate is called 'programming' or 'writing', and for reasons that will be explained later in this chapter, a programmed FGT is 'off' and shows a logical 0 when read; conversely, removing this stored charge from the floating gate is called 'erasing', and reading an erased FGT shows a logical 1.

Programming via hot carrier injection

Hot carrier injection (HCI) is a phenomenon in semiconductor devices where a carrier gains sufficient kinetic energy to overcome a potential barrier. During HCI of a FGT the carriers are accelerated by the potential difference between source and drain, and must overcome the potential barrier formed by the thin tunnel oxide layer. For the case of electrons as the hot carriers, this process is also referred to as hot electron injection (HEI) or channel hot electron (CHE) injection.

As shown in Fig. 3.8, the application of a positive drain voltage, V_{D} generates a lateral electric field, $\mathcal{E}_{\mathrm{lat}}$, accelerating the electrons towards the drain. Electrons that can overcome the tunnel oxide's Si/SiO_2 3.1eV energy barrier (ϕ_{B}) are injected into the floating gate from the high field pinch-off region near the drain, where the field is highest. The high field also induces impact ionisation, creating secondary electrons which can also be injected.

Without the added application of an 'elevated-on' control gate voltage, $V_{\mathrm{G}} >$

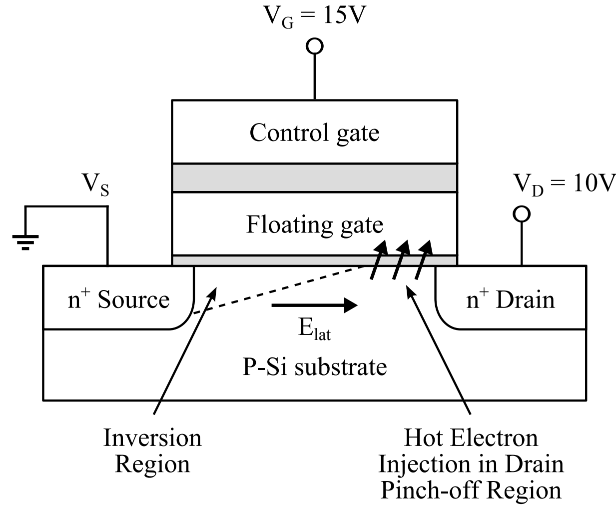


Figure 3.8: Simplified diagram of an FGT during programming by hot carrier injection with example drain and gate voltages shown.

V_D to attract them, the hot electrons injected into the tunnel oxide would return to the substrate channel. At the start of the process, the oxide electric field is attractive to electrons, allowing them to enter the floating gate. As the floating gate charges, the field eventually becomes repulsive to electrons, reducing the gate current to almost zero.

The energy band diagram for HCI is given in Fig. 3.9. As a generalisation, increasing the control gate voltage, V_G, increases charge on the floating gate, while increasing the drain voltage, V_D, increases the programming speed. The original FGT HCI process, no longer used due to inefficiency, is called drain-substrate avalanche; this process is similar to HEI, but hot holes are injected into the floating gate by applying a control gate voltage lower than the drain voltage ($V_G < V_D$).

One model describing the gate current as a result of HCI is the 'Lucky Electron Model' [9,10] originated by Shockley [11] and describes three events required for an electron to reach the floating gate. This process is outlined in Fig. 3.10.

First, an electron in the channel must gain sufficient energy from $\mathcal{E}_{lat}$ to become 'hot', its momentum must then be redirected towards the Si-SiO_2 interface

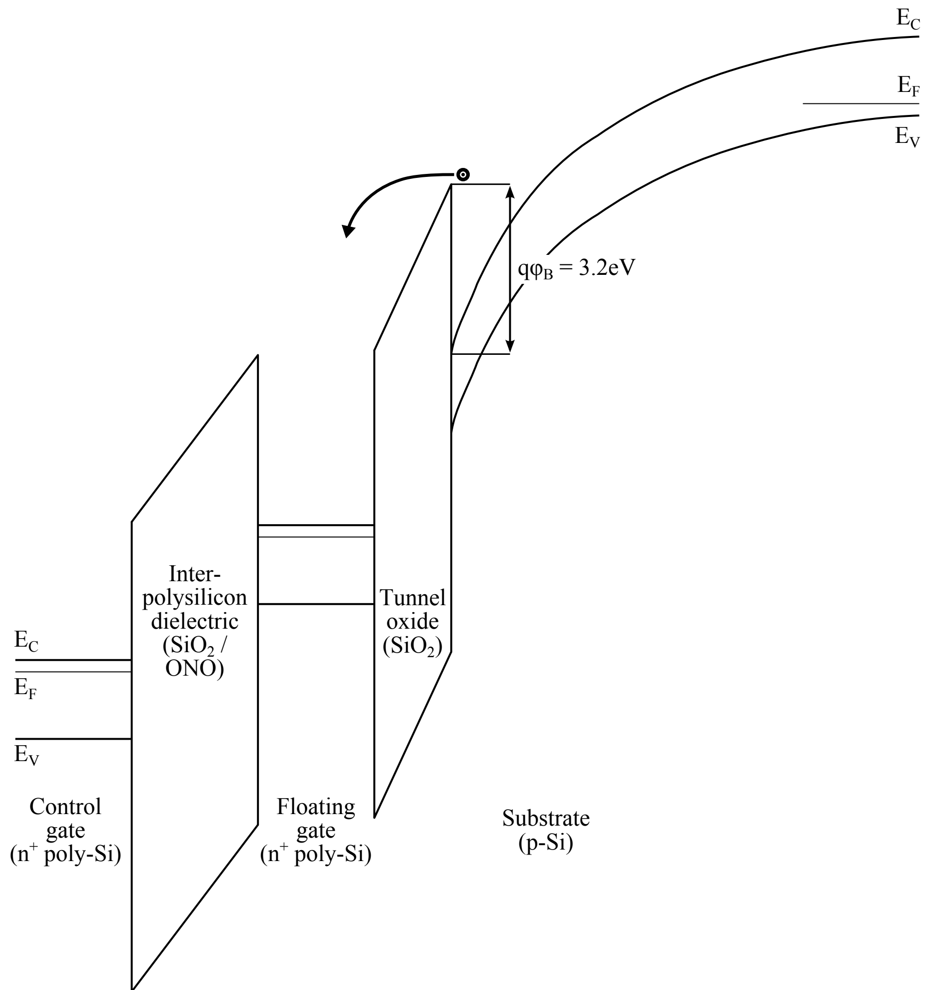


Figure 3.9: Energy band diagram of an FGT during programming by hot carrier injection.

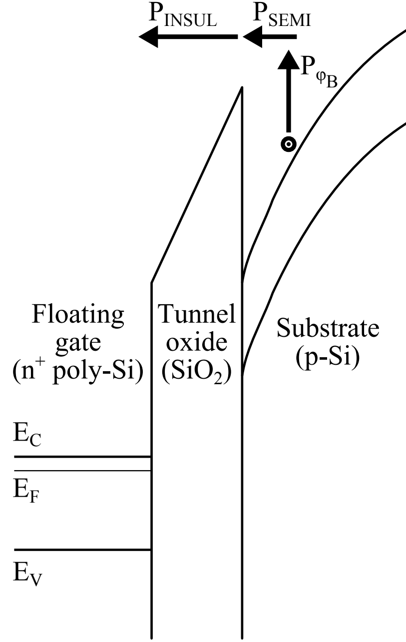


Figure 3.10: Energy band diagram showing the three processes in the 'Lucky Electron' model.

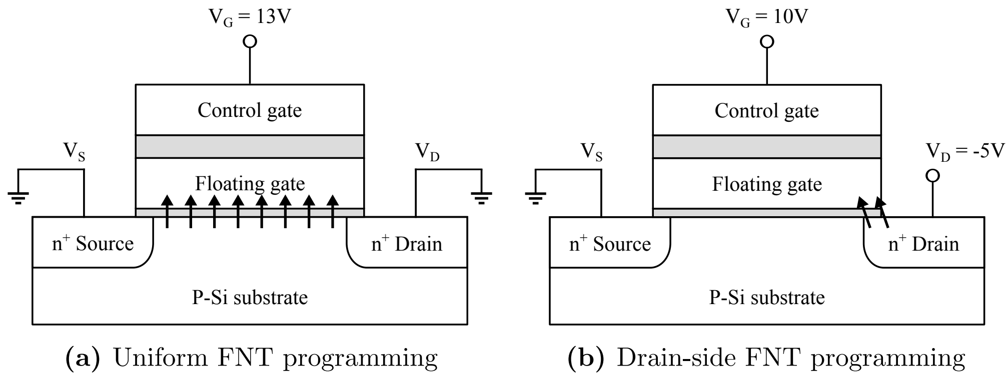
without suffering any momentum-altering collisions. The probability associated with this event, P_{ϕ_B}, is defined as the probability of an electron having sufficient normal momentum to overcome the Si-SiO_2 potential barrier. Second, the electron must suffer no collisions on its way to the tunnel oxide; the probability of this event occurring is denoted P_{SEMI}. Third, no collisions must impede the electron while travelling through the tunnel oxide; the probability of this is denoted P_{INSUL}. Since these three probabilities are statistically independent, the resulting probability is the product of each event. The gate current, I_G, is thus:

$$I_G = I_{DS} \int_0^{L_{eff}} \frac{P_{\phi_B} P_{SEMI} P_{INSUL}}{\lambda_r} dx \tag{3.1}$$

Where: λ_r is the mean free path of the momentum redirection scattering (92nm), L_{eff} is the effective channel length of the FGT, and I_{DS} is the drain-source current.

Programming via Fowler-Nordheim tunnelling

In modern EEPROM devices it is more common to use tunnelling rather than HCI as the programming mechanism. Applying a large positive voltage to the control gate generates an electric field which results in a triangular barrier, see Fig. 3.11a. FNT is the quantum-mechanical tunnelling of electrons through a triangular barrier; it differs from direct tunnelling in that only part of the thickness of the barrier is tunnelled through due to its shape. Electrons in the channel can tunnel through the partial width of the tunnel oxide barrier via FNT, and is sometimes referred to as 'tunnel injection'.



(a) Uniform FNT programming **(b)** Drain-side FNT programming

Figure 3.11: Simplified diagram of an FGT during programming by uniform **(a)** and drain-side **(b)** Fowler-Nordheim tunnelling with example voltages shown.

With an additional negative potential at the drain gate, it is possible to localise the FNT to only the drain-side region, shown in Fig. 3.11b. This alternative method is occasionally preferred over uniform FNT as it offers faster programming speeds due to an increased tunnelling current density as a result of confinement to a smaller injection area, however, this does put a greater strain on the drain-side tunnel oxide region than uniform tunnelling.

The tunnel oxide thickness, d_{ox}, for the transition between FNT and direct tunnelling can be approximated to:

$$d_{\mathrm{ox}} = \phi_{\mathrm{B}}/\mathcal{E}_{\mathrm{ox}} \tag{3.2}$$

Where: ϕ_B is the potential barrier height, and $\mathcal{E}_{ox}$ is the electric field at the tunnel oxide Si-SiO_2 interface. For ϕ_B=3.1V for electrons and $\mathcal{E}_{ox}$=6MV/cm for a medium tunnelling current, the transition oxide thickness at which direct tunnelling takes place is below ≈5nm [8].

Figure 3.12 is an energy band diagram of a FGT during FNT. In the diagram, E denotes an energy level, with subscript C and V referring to the conduction and valence bands respectively. For silicon, the energy band gap, $E_G = E_C - E_V$=1.12eV. The potential at the control gate generates a triangular energy barrier, ϕ_B, at the Si-SiO_2 interface. The shape of this barrier provides a shortened path for electrons in the substrate to tunnel through the thin tunnel oxide layer.

During FNT, the electric field across the tunnel oxide layer is critical; a positive gate voltage, V_G, creates electric fields in both the tunnel oxide and the IPD. From Gauss' law and the constant electric field we have [8, p. 353]:

$$\epsilon_{ox}\mathcal{E}_{ox} = \epsilon_{IPD}\mathcal{E}_{IPD} + Q_{FG} \tag{3.3}$$

$$V_G = V_{ox} + V_{IPD} = d_{ox}\mathcal{E}_{ox} + d_{IPD}\mathcal{E}_{IPD} \tag{3.4}$$

Where: ϵ is the permittivity, $\mathcal{E}$ is the electric field, d is the oxide thickness. The subscripts ox and IPD refer to the tunnel oxide layer and inter-polysilicon dielectric respectively, and Q_{FG} refers to the stored charge (negative for electrons) on the floating gate. Combining Eqs. (3.3) and (3.4) gives an expression for the tunnel oxide electric field dependence:

$$\mathcal{E}_{ox} = \frac{V_G}{d_{ox} + d_{IPD}(\epsilon_{ox}/\epsilon_{IPD})} + \frac{Q_{FG}}{\epsilon_{ox} + \epsilon_{IPD}(d_{ox}/d_{IPD})} \tag{3.5}$$

The tunnelling current density, J, during FNT is a function of the electric field, and takes the form:

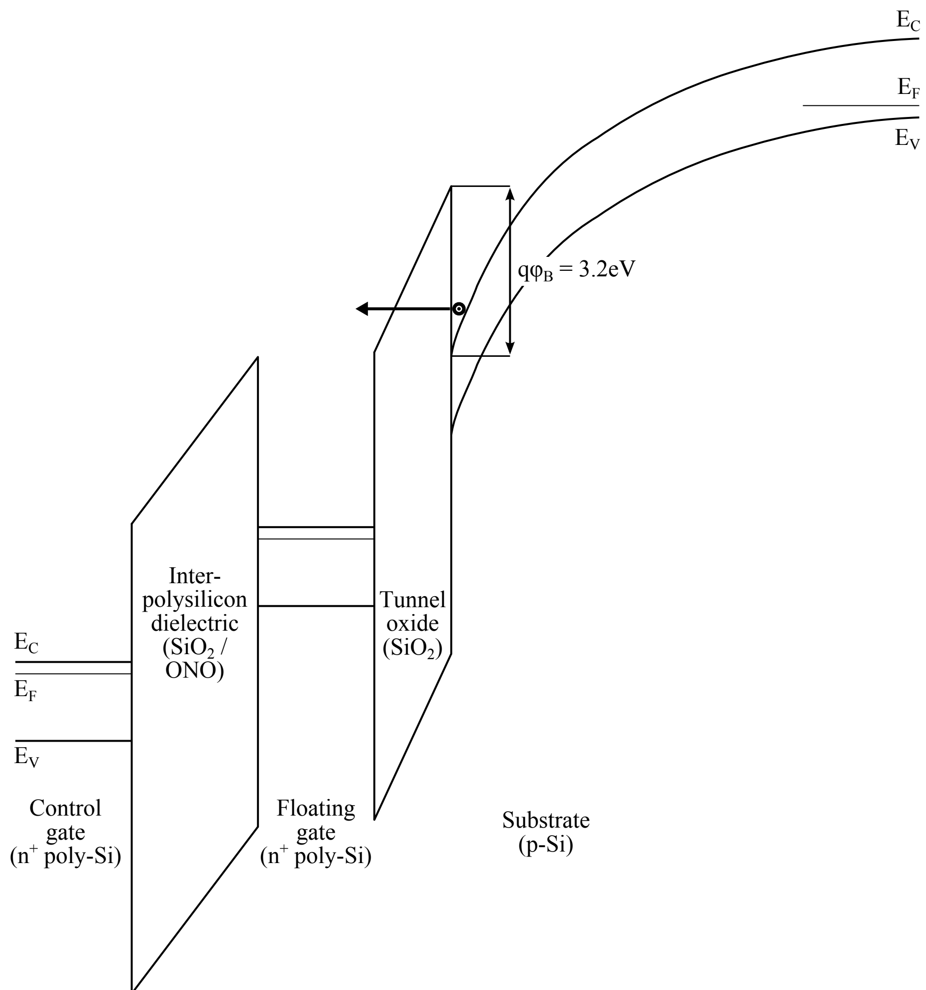


Figure 3.12: Energy band diagram of an FGT during programming by Fowler-Nordheim tunnelling.

$$J = \frac{e^2 \mathcal{E}_{\text{ox}}^2}{16\pi^2 \hbar \phi_{\text{B}}} \exp\left[\frac{-4\sqrt{2m_l^*}(e\phi_{\text{B}})^{3/2}}{3\hbar e \mathcal{E}_{\text{ox}}}\right] = A\mathcal{E}_{\text{ox}}^2 \exp\left(\frac{-B}{\mathcal{E}_{\text{ox}}}\right) \tag{3.6}$$

Where: A and B are constants in terms of effective mass and barrier height ($A = 9.63 \times 10^{-7}\text{A/V}^2$ and $B = 2.77 \times 10^8\text{V/cm}$ for thermal oxides), $\hbar = h/2\pi$ (h is the Planck constant), e is the unit of elementary charge, ϕ_{B} is the tunnel oxide potential barrier height, m_l^* is the longitudinal effective mass of an electron ($0.98 \times m_0$, m_0 is the rest mass of a free electron), and $\mathcal{E}_{\text{ox}}$ is the electric field at the tunnel oxide's Si-SiO_2 injecting surface, defined as:

$$\mathcal{E}_{\text{ox}} = \frac{V_{\text{ox}} - V_{\text{FB}}}{d_{\text{ox}}} \tag{3.7}$$

Where: V_{ox} is the potential across the tunnel oxide, and V_{FB} is the flat band voltage.

Erasing via Fowler-Nordheim tunnelling

Erasing cells of charge using FNT is carried out in much the same way as programming except with inverted potentials applied, it is sometimes known as 'tunnel release'. Similar to programming, the control gate potential generates an electric field resulting in a triangular potential barrier. This triangular shaped barrier provides a path for electrons to tunnel through the tunnel oxide layer, shown in Fig. 3.13.

By applying a large negative potential to the control gate and keeping the source and drain grounded, as shown in Fig. 3.14a, electrons uniformly tunnel from the floating gate, through the tunnel oxide, to the substrate – erasing the floating gate of stored charge. By additionally applying a positive drain voltage, tunnelling can be confined to the drain region. As previously mentioned, uniform tunnelling is slower than drain-side tunnelling (shown in Fig. 3.14b), however, drain-side tunnelling causes a greater degree of tunnel oxide damage in the drain region since a small area is bombarded by a much higher current density.

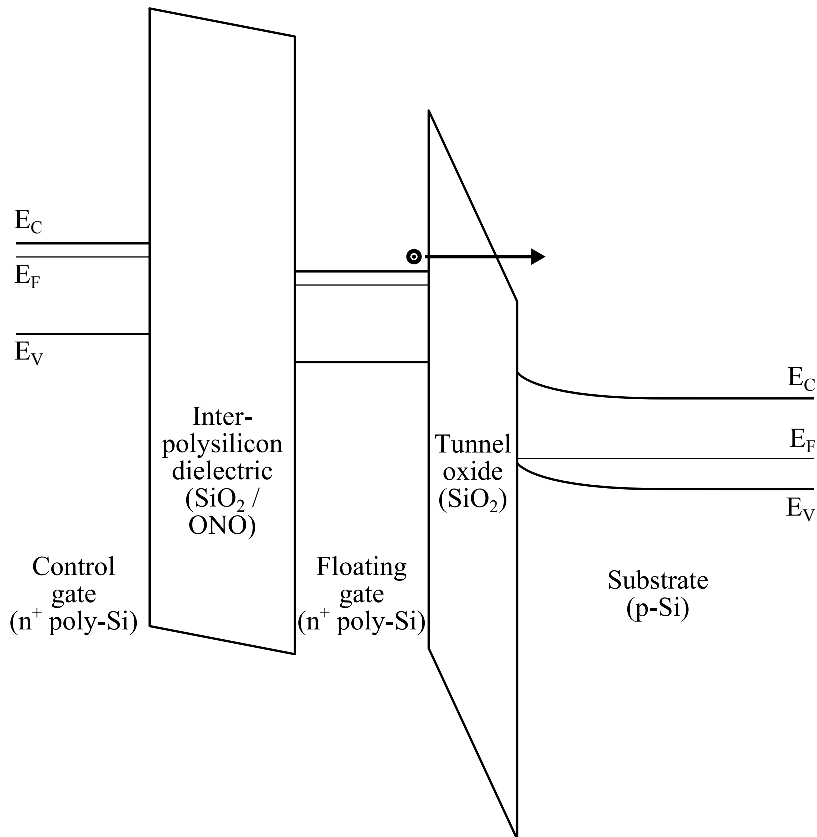
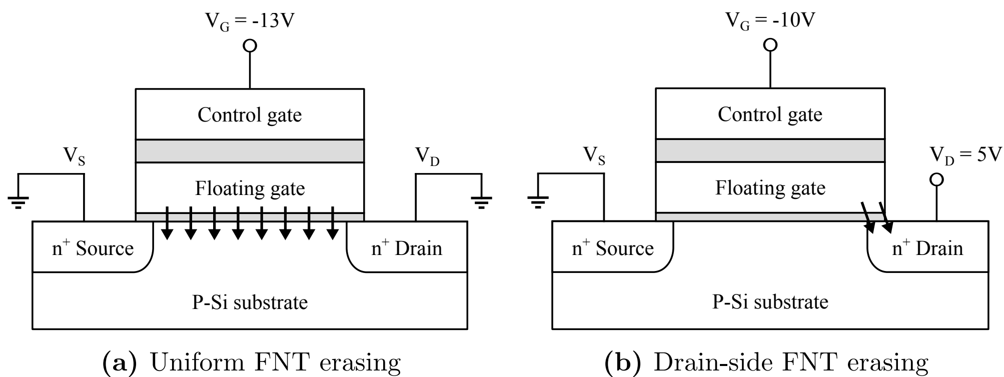


Figure 3.13: Energy band diagram of an FGT during erasing by Fowler-Nordheim tunnelling.



(a) Uniform FNT erasing **(b)** Drain-side FNT erasing

Figure 3.14: Simplified diagram of an FGT during erasing by uniform **(a)** and drain-side **(b)** Fowler-Nordheim tunnelling with example voltages shown.

Reading a flash cell

The physical mechanisms described previously act to alter the amount of charge stored within the floating gate. The charge stored within the floating gate varies according to:

$$\Delta Q_{\mathrm{FG}} = I_{\mathrm{G}} \Delta t \tag{3.8}$$

Where: ΔQ_{FG} is the change in charge within the floating gate, I_{G} is the gate current, and Δt is the programming time. Injecting charge into the floating gate alters the threshold voltage, V_{T}, of the FGT by:

$$\Delta V_{\mathrm{T}} = \frac{-\Delta Q_{\mathrm{FG}}}{C_{\mathrm{FG}}} \tag{3.9}$$

Where: C_{FG} is the sum of the capacitances across the IPD and tunnel oxide layers surrounding the floating gate. This equation gives a range between the threshold voltages of an erased (natural state) cell and one that has been programmed with charge, otherwise known as the logic margin window. These injected charges screen the control gate's electric field, thus modifying the threshold voltage of the transistor:

$$V_{\mathrm{T}} = V_{\mathrm{T(erased)}} - \frac{\Delta Q_{\mathrm{FG}}}{C_{\mathrm{FG}}} \tag{3.10}$$

Where: $V_{\mathrm{T(erased)}}$ is the natural (erased) threshold voltage of the FGT, and Q_{FG} is negative for the case of electrons, thus shifting the threshold voltage positively. The shift in the FGT threshold voltage affects the conductance of the channel – the threshold voltage can also be defined as:

$$\Delta V_{\mathrm{T}} = -\frac{d_{\mathrm{IPD}} \Delta Q_{\mathrm{FG}}}{\epsilon_{\mathrm{IPD}}} \tag{3.11}$$

This voltage shift can be directly measured in a plot of gate voltage (V_{G})

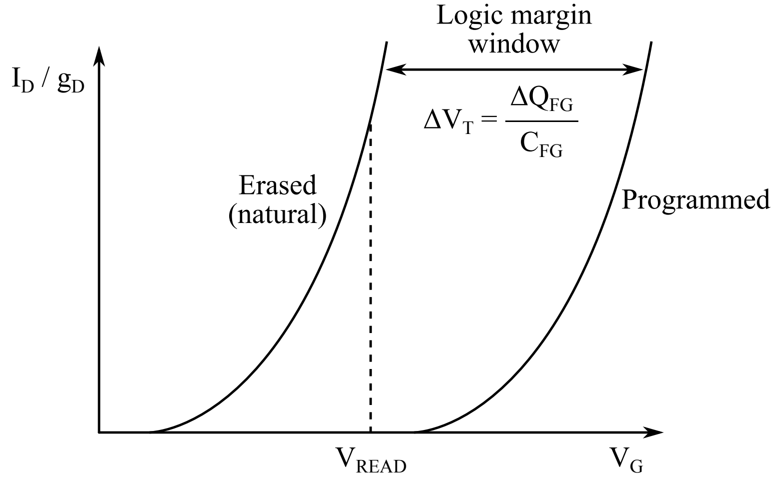


Figure 3.15: $I_{\mathrm{DS}} - V_{\mathrm{CG}}$ transfer characteristics of an FGT cell. The 'logic margin window' is the difference between the threshold voltages of the erased (natural) state and the programmed state. During reading, a voltage, V_{READ}, within this window is placed on the control gate to determine the logical state of the cell.

against either the drain current (I_{D}) or the drain conductance (g_{D}). For small drain voltages, the channel conductance for $V_{\mathrm{G}} > V_{\mathrm{T}}$ in an n-channel MOSFET is given by:

$$g_{\mathrm{D}} = \frac{I_{\mathrm{D}}}{V_{\mathrm{D}}} = \frac{Z}{L}\mu C_{\mathrm{FG}}(V_{\mathrm{G}} - V_{\mathrm{T}}) \tag{3.12}$$

Where: Z is the channel width, L is the channel length, μ is the charge carrier effective mobility, and C_{FG} is the sum of the IPD and tunnel oxide capacitance per unit area.

An erased FGT has a (natural) threshold voltage, $V_{\mathrm{T(erased)}}$, while a programmed FGT has a more positive threshold voltage, $V_{\mathrm{T(programmed)}}$. The range of voltages between $V_{\mathrm{T(erased)}}$ and $V_{\mathrm{T(programmed)}}$ is known as a memory's 'logic margin window' (see Fig. 3.15).

To read a cell, a voltage within the logic margin window, V_{READ}, is applied to the control gate and a low voltage is applied to the drain, see Fig. 3.16. The cell will either form a conductive channel between source and drain (logical 1), or remain insulating (logical 0), depending on whether or not the electric field

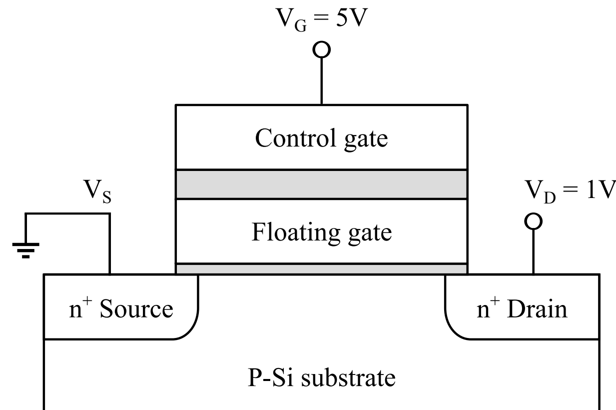


Figure 3.16: Simplified diagram of an FGT during reading with example low drain voltage and V_{READ} gate voltage shown.

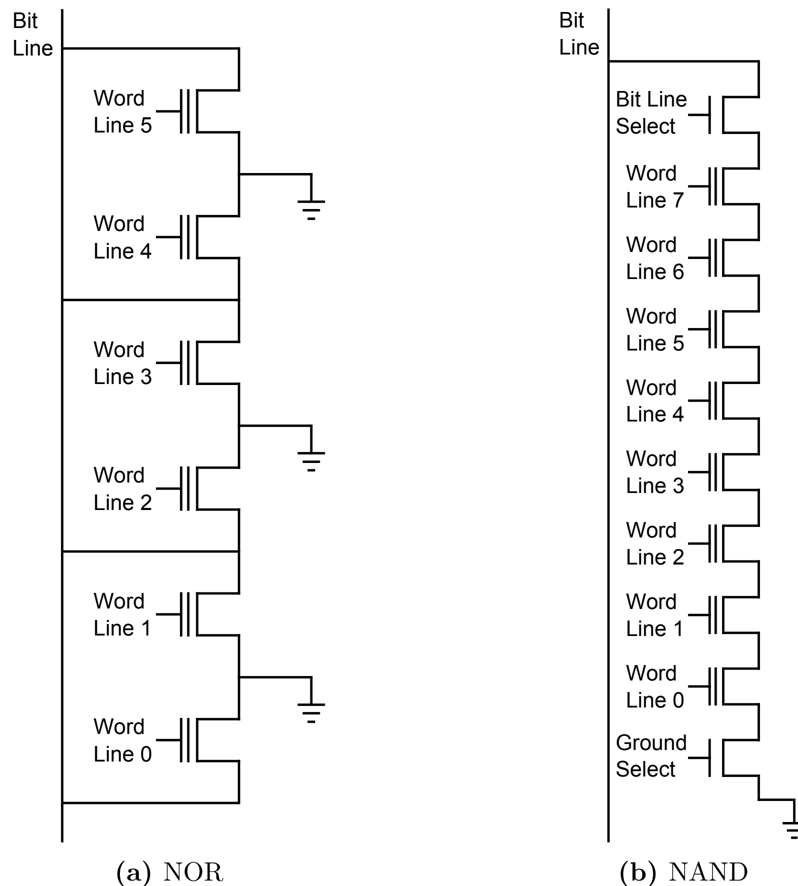
generated by the control gate has been screened by the charge within the floating gate.

The returned output from reading the cell is thus: a FGT that has been programmed with charge will screen the control gate voltage and return a logical 0 output; conversely, an erased FGT returns a logical 1 output. As mentioned previously, the 'page' erasing system in flash memory means that logical 0 (programming) can be written as single bits, bytes, or words; however, logical 1 (erasing) can only be written block-wise.

3.4.3 NOR and NAND configuration

The two main architectures that exist for flash memory are: NOR and NAND, although others do exist *e.g.* AND and DINOR (divided bit-line NOR). The first to be introduced was NOR memory, developed by Intel and released in 1988. To reduce the cell area of NOR memory, which requires 1 contact per 2 cells, NAND was developed, and was introduced by Toshiba in 1989. NAND allows for a substantial reduction in ground wires and bit lines, which consume most of the area in a memory array.

In NOR memory, Fig. 3.17a, each FGT is connected to ground and the bit line, with one transistor per memory cell. This design leads to the cells



(a) NOR **(b)** NAND

Figure 3.17: NOR-configuration **(a)** and NAND-configuration **(b)** memory architecture.

resembling NOR logic gates: when the word line (control gate input) is brought high (above V_T), the storage transistor in the cell pulls the output bitline (drain output) low. To program charge into the floating gates NOR typically uses CHE processes, while removing stored charge is carried out by FNT. In general, NOR reads faster and can perform random-access tasks, but is slower at programming and erasing operations than NAND. Due to its lower density, it is also more expensive than NAND.

In NAND memory, Fig. 3.17b, the cells behave in a similar way to logical NAND gates – only if all word lines are brought high is the bit line pulled low. To read a selected NAND flash cell, each of the word lines (control gates) in the NAND series is pulled up above the threshold voltage of a programmed FGT

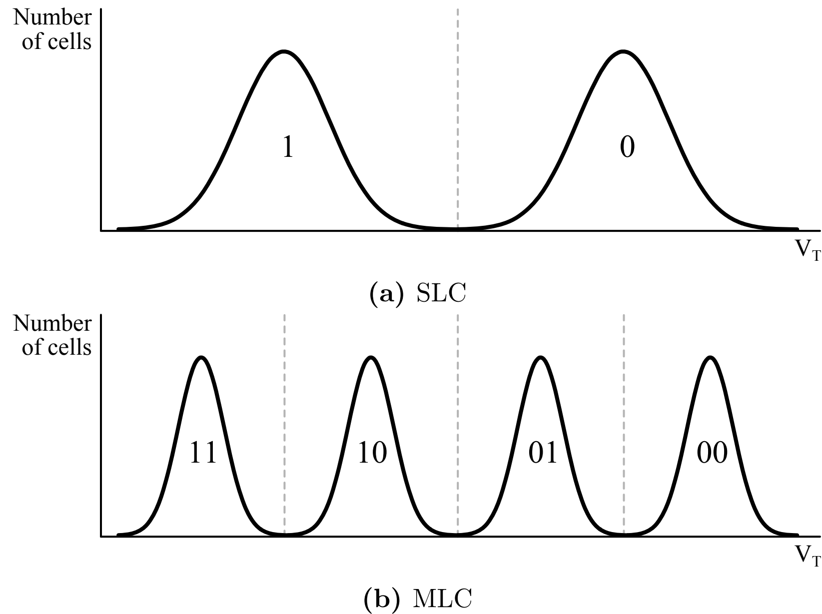
($V_{T(programmed)}$), while the selected cell is pulled up to just above the threshold voltage of an erased FGT ($V_{T(erased)}$). The series of FGTs will only conduct if the selected FGT is in the erased state, *i.e.* no stored charge (logical 1), but remains insulating if it has stored charge (logical 0). NAND memory typically uses FNT for both write and erase operations.

Each series of FGTs is connected in a NOR-style bit line array via some additional transistors, controlling access to individual NAND series. Even taking into account the additional transistors controlling access, NAND has a much higher transistor density than NOR due to the decrease in ground wires and bit lines. The rise in larger applications, photos, games, multimedia, and the need for mobile mass-storage has put NAND ahead of NOR in many markets, including solid state memory chips for the mobile phone handset market.

When reading a NAND cell, a weaker signal is detected by the sense amplifier, compared to NOR, due to multiple transistors in series. This has the effect of reducing the read time, but this issue can be rectified by operating in a serial access mode. It also requires RAM to execute program code, *i.e.* it cannot perform 'execute-in-place' (XIP) like NOR memory. However, NAND has the advantage of being able to more quickly erase blocks and write new data, with write and erase power consumption far reduced compared to NOR.

3.4.4 Multi-level cell memories

Flash memory originally consisted only of single-level cell (SLC) devices where one bit was stored per cell, *i.e.* the presence of charge on the floating gate was simply a binary condition. Multi-level cell (MLC) devices store multiple bits per cell by more accurately controlling the level of charge injected into the cell and through the use of incorporated current flow sensors to measure this level of trapped charge. A comparison of the typical threshold voltage distributions of SLC and 2 bit (four state) MLC memory is given in Fig. 3.18 and portrays the correspondingly tighter logic margin window found in MLC memories.



(a) SLC

(b) MLC

Figure 3.18: Diagrams of example single-level cell (SLC) **(a)** and 2-bit multi-level cell (MLC) **(b)** memory voltage distributions.

The number of individual states required is 2^n, where n is the number of bits per cell, *e.g.* to identify the 4 bits of data stored in a single 4-bit MLC (as found in modern high-end NAND flash memories) requires 16 distinct logical states. This reduces the available logic margin window size considerably, and the 2009 ITRS Executive Summary stated explicitly that near-future high density MLC cells would have less than 10 electrons separating individual threshold voltage windows. The reduced storage capacity of SLC memories make them a costlier solution, but due to the larger logic margin window are capable of offering faster transfer speeds and a greater endurance.

3.4.5 Memory reliability

At this point it is worth clarifying the difference between endurance and retention. Endurance defines the number of programming cycles a memory can withstand. Retention is a measure of the time a memory cell can retain charge and still differentiate between logical states. Both the endurance and retention are heavily dependent upon the quality of the tunnel oxide and IPD layers.

Oxide breakdown, when the oxide is destroyed and becomes conductive, occurs after a fixed amount of charge, Q_{B}, has been injected, and is a function of the applied electric field [12]. This measurement is used as an industry standard test to determine a MOSFET's gate oxide quality – the aim is to achieve a high value of Q_{B}.

'Trap-up' is defined as the trapping of mobile, positively-charged holes in the oxide during programming operations and is the primary failure mechanism for tunnel oxides. Oxide defects and broken bonds within the thin oxide layer serve as trapping centres for these holes [13]. These trapped charges modify the electric fields during injection, altering the amount of charge that can be transferred into and out of the floating gate.

The purpose of the IPD layer is to isolate the floating gate from the electrodes (source, drain, control gate, and substrate). Since the floating gate is polycrystalline silicon, it is typically oxidised during IPD growth processes. This oxidation changes the surface topography and leads to the formation of surface asperities at the interface due to enhanced oxidation at grain boundaries [14]; these features cause electric field enhancement which result in higher leakage currents. The IPD quality is also determined by a variety of other factors, such as gate doping level and the temperature at which gate deposition takes place during fabrication [15].

As mentioned earlier, it is now commonplace to use an ONO composite stack to form the IPD. These multiple dielectric stacks have the advantage of a lower defect density and better electric field properties [16]. Any electrons that leak from the floating gate become trapped in the oxide-nitride interface, building up an electric field which acts to oppose further charge leakage from taking place, resulting in a lower leakage current.

Endurance

The logic margin window, as previously defined, is the difference between the threshold voltages of the programmed and erased states. Regular operation causes damage to the tunnel oxide, with non-uniform operations such as HCI being the main culprits. After repeated programming cycles, the tunnel oxide begins to degrade due to injected electrons becoming trapped, reducing the ability of the memory cell to store charge. This causes the logic margin window to close up, see Fig. 3.19, and eventually the two states will become indistinguishable from one another. The endurance is defined as the number of write-erase cycles at which logical 1 and logical 0 can no longer be distinguished from one another.

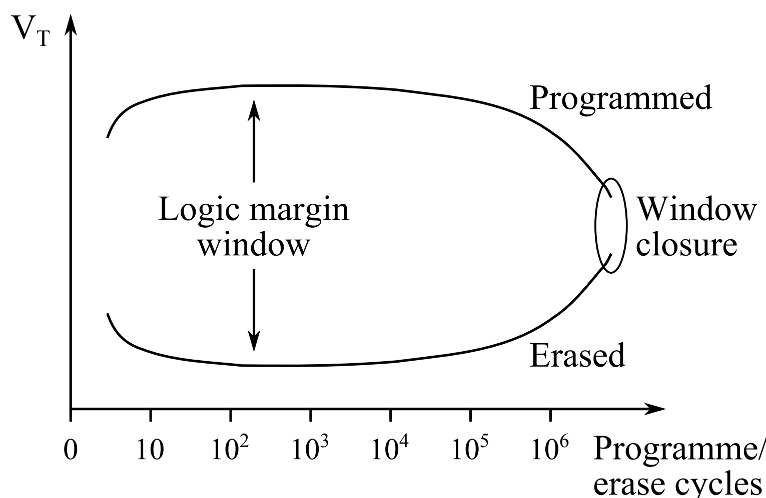


Figure 3.19: A plot of the endurance of an EEPROM/flash cell showing the typical logic margin window variation as a function of the number of programme/erase cycles.

Most commercially available SLC flash memories are guaranteed to withstand a minimum of 10^5 program/erase cycles, some up to 10^6. By comparison, due to tighter threshold voltage windows, 4-bit MLC are typically only capable of 10^4 P/E cycles, and 8-bit cells only 5000 P/E cycles. To improve endurance, high quality tunnel oxides are required to endure the high electric field stresses from programming/erasing.

Retention

The charge stored within the floating gate can leak away through the surrounding dielectric layers – the tunnel oxide and the IPD. In the same way that increasing the charge increases the threshold voltage of the cell, see Eq. (3.9), charge leakage reduces it. Charge can be lost through numerous methods [17, 18] such as thermionic emission, electron detrapping, and mobile ion contamination. The number of electrons lost through leakage is simply:

$$\text{Number of electrons} = \frac{\Delta Q_{\text{FG}}}{e} \tag{3.13}$$

Substituting in Eq. (3.8), the number of electrons lost is:

$$\text{Number of electrons} = \frac{I_{\text{G}} \Delta t}{e} \tag{3.14}$$

As a simple example: for a cell with a C_{FG} of 30fF and V_{T} of 3V, the charge stored (ΔQ_{FG}) is 9×10^{-14}C. This translates as 5.625×10^{5} electrons – to remove this many electrons would drop the V_{T} by 3V and erase the cell. So we can calculate the retention time, Δt, as a function of leaked gate current, I_{G} – see Table 3.3. From this we can see that it would take a leakage current of 2.85×10^{-22}A a period of 10 years to shift the threshold voltage by 3V through charge leakage. Improving the quality of the IPD and tunnel oxide layer acts to lower the leakage current experienced by a cell, preserving written data for longer.

However, there are other factors to consider. For SLC memory, a 50% drop in voltage is sufficient to swap the state of a cell to a logical 1. Also, as electrons leak out of the floating gate, further leakage becomes less likely due to a decreased repulsion between the remaining stored charges. The retention time, t_{R}, is the amount of time taken to decrease the stored charge within a floating gate to 50% of its initial value [8, eq. 122, p. 357]. This is calculated by:

Table 3.3: Table showing retention time (Δt) as a function of leaked gate current (I_G).

Leakage Current (I_G/A)	Retention Time (Δt/Years)
1×10^{-20}	0.29
5×10^{-21}	0.57
1×10^{-21}	2.8
5×10^{-22}	5.7
2.85×10^{-22}	10

$$t_\mathrm{R} = \frac{\ln 2}{\nu \exp(e\phi_\mathrm{B}/kT)} \tag{3.15}$$

Where: ν is the dielectric relaxation frequency; e is unit of elementary charge; ϕ_B is the height of the potential barrier to be overcome; k is the Boltzmann constant; and T is the absolute temperature. From this equation we can see clearly the high dependence of retention time on temperature.

As can be seen in Fig. 3.20 from Ielmini *et al.* [19], a tunnel oxide thickness of 4.3nm (into the region of direct tunnelling, as opposed to FNT, for poly-Si floating gate material) is required to meet the 10 year minimum requirement for flash memory data retention time. This is a simple prediction and fails to take into account that a full erase is not required to change the logical state of a cell, or other considerations such as defect-assisted tunnelling arising after multiple P/E cycles. A better estimation is the time to remove 20% of the stored charge, which results in a 4.5nm tunnel oxide losing 20% charge in 4.4 minutes, 5nm in 1 day and 6nm in 0.5–6 years. From this, <7nm tunnel oxide thickness offers insufficient retention times, making 7–8nm tunnel oxide thickness a minimum requirement. Data sheets often specify data retention to exceed 100 years at 25°C and this higher minimum value has been accepted by all vendors across the industry to be sufficient given error margins in fabrication.

Aritome *et al.* [20] tested and compared the data retention characteristics of EEPROM memory with bipolarity FNT write/erase technology with devices

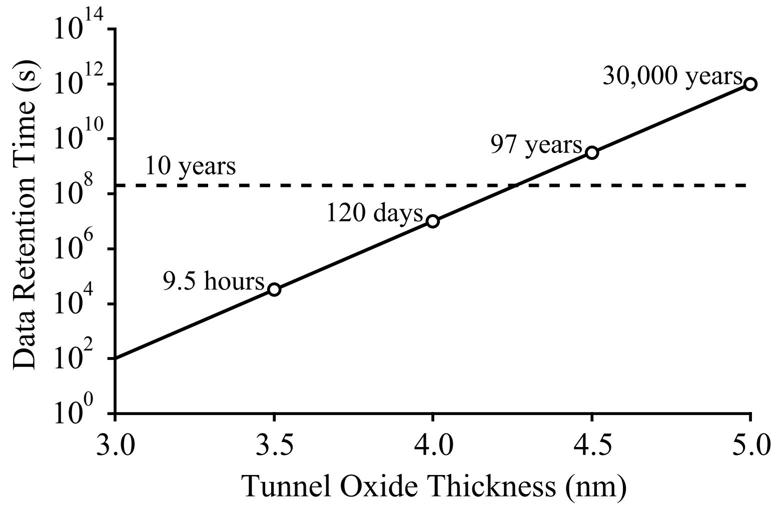


Figure 3.20: Calculated data retention times of flash memories as a variant of tunnel oxide thickness. After: [19].

utilising CHE injection write and FNT erase technology. For devices of ≈10nm tunnel oxide thickness these two technologies were found to offer similar retention times, however, for devices with <9nm tunnel oxide thickness the bipolarity FNT technology offers a superior data retention time due to lower stress-induced oxide leakage currents. At 7.5nm tunnel oxide thickness the retention time was approximately 50 times greater. This bipolarity FNT write/erase technology has facilitated the scaling down of the tunnel oxide thickness below 9nm, which in turn allowed for lower operating voltages and faster read operations.

Devices known as SONOS (Silicon-Oxide-Nitride-Oxide-Silicon) are under development as a possible successor to flash memory devices. Using silicon nitride (Si_3N_4) instead of poly-Si as the floating gate material allows for thinner tunnel oxide and blocking oxide (equivalent of IPD for SONOS devices) layers due to a greater homogeneity. These devices may offer higher endurance and lower programming voltages over traditional poly-Si flash memories. SONOS devices degrade differently from EEPROM devices, electrons may become trapped permanently in the oxide layer, but this is rare. SONOS devices are capable of 10^7 P/E cycles. The tunnel oxide thickness is 2nm, with a scaling limit of just 7Å, and blocking oxide of 5.5nm.

Accelerated retention testing

Testing the number of P/E cycles a memory device can perform is a relatively quick test to conduct; at 1 cycle per second (a very low estimate with current technology), 10^6 cycles can be tested in under 12 days. It is also possible to simulate extended exposure of the chip to detrimental conditions by increasing the level of exposure and calculating the acceleration factor (AF). Tests models such as contact corrosion, humidity, magnetic field, electrostatic discharge, X-ray and UV exposure have all been developed and standardised[6]. From such acceleration models, tests can be conducted to simulate many years of common exposure/use in a matter of hours, see Eq. (3.16). The acceleration factor (AF) relates the amount of time the device is under test, t_{test}, to the real world application time this relates to, t_{use}.

$$t_{\text{test}} = \frac{t_{\text{use}}}{\text{AF}} \tag{3.16}$$

It is worth noting that MLC memory, due it having multiple bits per cell, has a lower maximum operating temperature than SLC memory. As the temperature increases, the leakage current increases; with tighter constraints on logic margin windows, charge loss is more significant to data loss in MLC devices than their SLC variants. Also worth noting is that for charge leakage to occur, the floating gate must initially hold charge, *i.e.* it must first be programmed into a logical 0 state. Thus, single bit failures caused by leakage currents draining the stored charge only read erroneous logical 1 states, never the other way round.

One of the most important tests to this investigation in particular is the data retention test. Memory cells are typically guaranteed to retain data for 10 years, and are commonly stated to retain data for hundreds, even thousands of years. Obviously testing for this entire period is unfeasible, thus an acceleration model

[6] A short and non-exhaustive summary can be found in [21, Table 9.5], and also online at: siliconfareast.com/reltests.htm and subsequent pages

known as a 'stabilisation bake'[7] is used, see Eq. (3.17) [21, 22]. The leakage-induced data retention temperature dependence is an Arrhenius equation – a simple and highly accurate formula describing the temperature dependence of reaction rates. The memory is written with a sample data set prior to testing, and examination of the memory after testing will give an indication of the reliability.

$$\text{Acceleration Factor} \qquad \mathrm{AF} = \exp\left[\frac{E_\mathrm{a}}{k}\left(\frac{1}{T_1} - \frac{1}{T_2}\right)\right] \tag{3.17}$$

Where: E_a is the activation energy (0.6eV); k is the Boltzmann constant (8.617×10^{-5}eVK^{-1}); and T_1 and T_2 are test and application temperatures respectively.

Inputting T_2 =298K (25°C), and conducting a stabilisation bake test at T_1 =323K (50°C) gives AF≈6. This acceleration factor of 6 means that the data 'ages' 6× faster, *i.e.* for every hour at 50°C the data ages as if had been kept for 6 hours at an application temperature of 25°C.

This also means that with a device's data retention lifetime known beforehand, the data retention time at a specific temperature can be calculated. For example, a device with a 100 year data lifetime at 25°C has only a 16.67 year lifetime at 50°C. This can also be used to model variable temperature applications, *e.g.* with known proportions of time the device is held at different working temperatures.

Venkat and Haensel [23] at Texas Instruments conducted a study of an MSP430 flash memory device. First calculating the data lifetime, then adjusting this figure through successive stabilisation bakes at temperatures up to 250°C and feeding it back into the equation to more accurately quantify its true data retention capabilities. The results finally concluded that the MSP430 is capable of retaining data at 25°C for ≈1315 years. This figure was subsequently adjusted

[7] The stabilisation bake methodology is outlined in detail within MIL-STD 883/1008

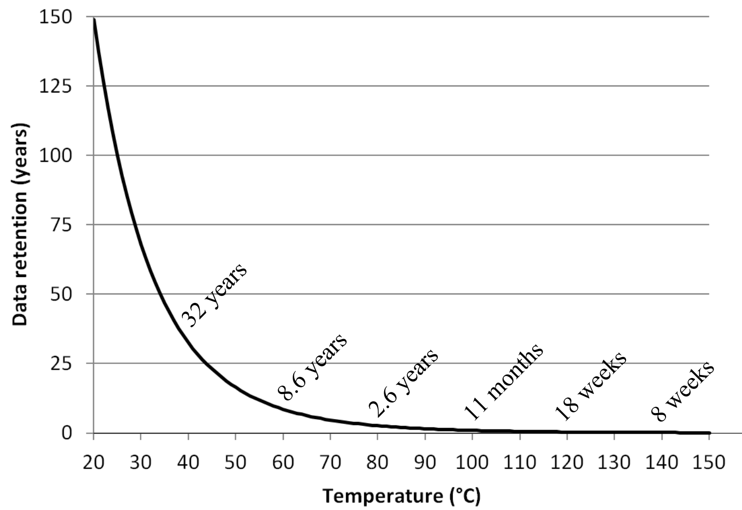


Figure 3.21: Data retention graph for flash memory with an industry standard retention of 100 years 25°C during operation at temperatures up to 150°C.

slightly to ≈1324 years in a later application note published on the MSP430 by Forstner [22] also at Texas Instruments.

Therefore, in a constant temperature application at 50°C, the data retention time of an MSP430 device is approximately 220 years; at 100°C the data has a lifetime of 12 years; at 160°C, 11 months; and at 400°C, 1 day. Using this method it is possible to calculate the data aging that would take place should the sample be subjected to extreme heat, such as in a house fire or evidence disposal scenario; see Fig. 3.21 for the Arrhenius plot of a device with an industry standard 100 year retention 25°C.

3.5 Fire investigation study

One important study that should be mentioned at this point is the *Full-Scale House Fire Experiment* [24]. This is a test report of an investigation carried out by the Building and Fire Research Laboratory[8]. This investigation of a staged fire experiment in a single-family house took recommendations from fire

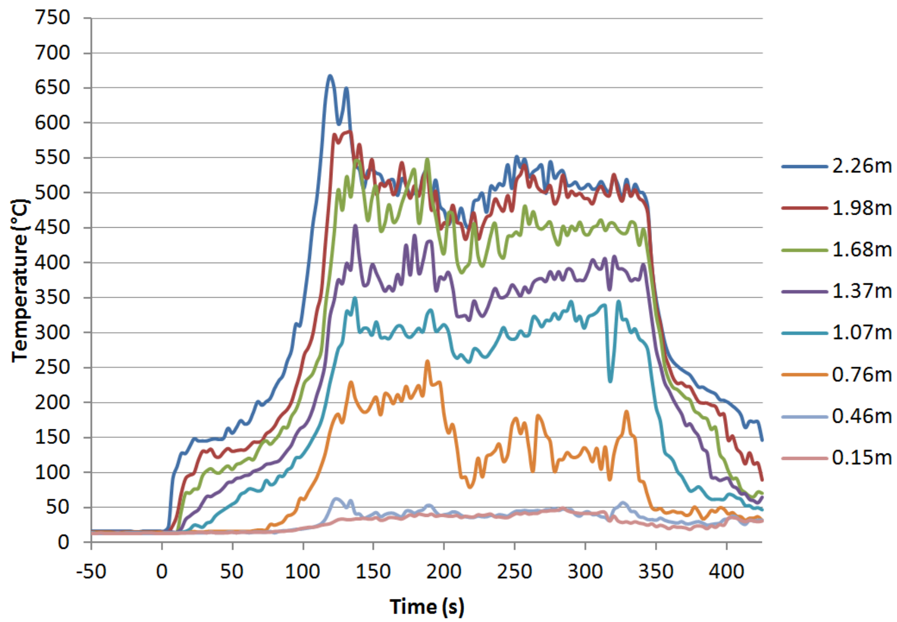
[8]part of the National Institute of Standards and Technology (NIST)

models and empirical correlations. The temperature and radiant heat flux was monitored from numerous sensors throughout the building. The fuel consisted of 1 litre of two-cycle engine fuel and household furnishings.

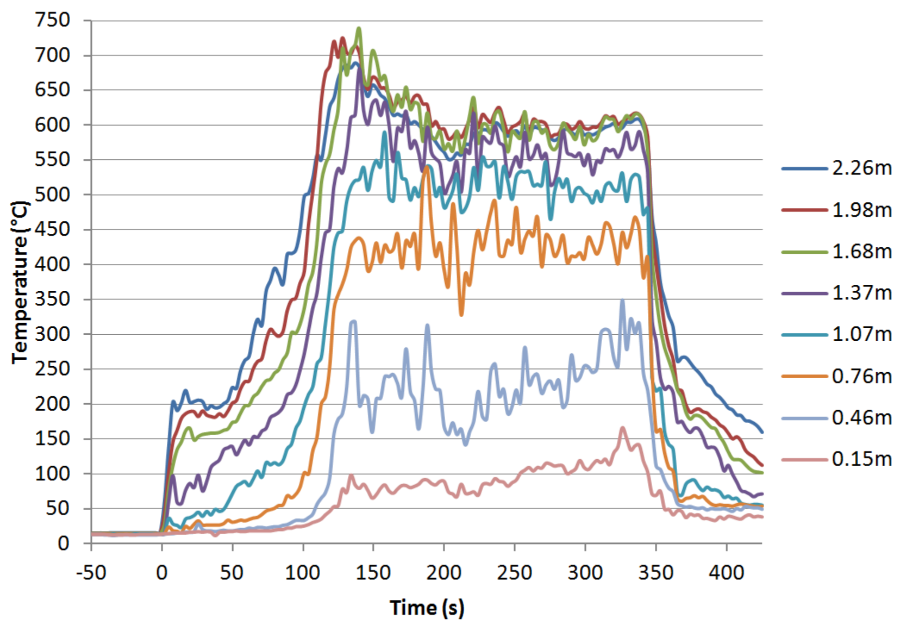
This investigation offers a valuable insight as to the typical temperatures that a SIM card may be subjected to, should it be recovered from a house-fire. Bare-bead type-K thermocouple sensors were mounted every 30cm up to ceiling level in both the dining room and living room, and temperatures recorded every 3 seconds for 7m30s, with ignition occurring at 50s.

The peak temperatures achieved in the dining room, see Fig. 3.22a, were $\approx$680°C, however, this peaked only for a short period of time and levelled out at $\approx$510°C shortly afterwards for the remainder of the test. The living room temperatures, Fig. 3.22b, were higher, peaking at $\approx$740°C and stabilising at $\approx$600°C, but also the average temperature was notably higher in the living room, with 1m height sensors recording $\approx$200°C higher sustained temperature; this could be accounted for with the different furnishings present in the living room.

Since floor or desk height is generally where a mobile phone will reside, the maximum peak temperatures expected from a house fire will be $\approx$175°C (while not truly floor-height, this is the maximum recorded temperature at 15cm) and $\approx$540°C respectively. The maximum sustained temperatures at floor and desk levels should be $\approx$80°C and $\approx$420°C respectively. The acceleration factors and data retention times for these maximum sustained and peak temperatures are shown in Table 3.4. It is clear that data stored within a SIM card could potentially survive a house fire environment at desk or floor height, however, at desk height peak temperatures of 540°C this data will only survive for around 20 minutes for devices with an industry standard 100 year retention at 25°C.



(a) Dining room



(b) Living room

Figure 3.22: Dining room **(a)** and living room **(b)** thermocouple temperatures (legend shows height of different thermocouple placement) from a full-scale house fire test. After: [24].

Table 3.4: Acceleration factors and data retention times for 100 year flash memory industry standard and acceleration-tested 1324 year MSP430 flash memory lifetimes for house fire temperatures at desk and floor height. The 3 letters next to each application temperature figures designate **D**ining / **L**iving room, **d**esk / **f**loor height, **p**eak / **s**ustained temperature.

Constant temperature application (°C)	Acceleration factor (AF) of a 25°C application	Approximate retention time	
		Industry standard 100 years 25°C	MSP430 device 1324 years 25°C
35 (Dfs)	2.13×10^{0}	47 years	620 years
50 (Dfp)	6.09×10^{0}	16 years	220 years
80 (Lfs)	3.8×10^{1}	2.5 years	35 years
140 (Dds)	6.65×10^{2}	8 weeks	2 years
175 (Lfp)	2.48×10^{3}	2 weeks	6 months
260 (Ddp)	2.96×10^{4}	30 hours	16 days
420 (Lds)	6.02×10^{5}	90 minutes	19 hours
540 (Ldp)	2.65×10^{6}	20 minutes	4 hours

3.6 Forensic evidence within SIM cards

The files stored in a smart card are protected from access through use of a hierarchical file system. This system consists of 3 distinct file types: the top-most level is termed a Master File, beneath this may be several levels of Dedicated Files and finally the data stored within Elementary Files. Access to an Elementary File must pass through each of its parent files in turn. The Dedicated Files contain access rules for the layer beneath them, thus the logical access channel to the Elementary File is formed and access is controlled.

In disposing of a SIM card/mobile phone, a user seeks to eliminate the link between them and the activities recorded on the SIM/phone/smart card. Obtaining and associating the discarded, often damaged, SIM/phone with the user allows a forensic examiner to build up a picture of their involvement. When carrying out examination of electronic devices it is common to conduct this interrogation within Faraday cages to block incoming and outgoing signals which could overwrite data within the device, thus preserving the integrity of evidence.

An 8 kilobyte SIM card has the capacity to store 100 telephone numbers and

their corresponding names, 15 recent SMS messages, PIN information and 25 most-recent incoming/outgoing numbers on the call register. For higher-end SIM cards the EEPROM can be replaced by a flash array and could potentially store megabytes worth of numbers, messages, photos, videos, emails, etc. Evidence can be found in various files on a SIM card, including most notably the following:

- Integrated Circuit Card ID (ICC-ID): Unique to that specific SIM card, this number can be up to 19 digits long and contains a single check digit calculated using the Luhn algorithm [25];
- International Mobile Subscriber Identity (IMSI): This SIM-specific identity number consists of a maximum of 15 digits and uniquely identifies the user to the network when switching on the phone or coming within range of a mobile base station. The IMSI consists of three parts: a 3 digit Mobile Country Code (MCC), a 2 digit Mobile Network Code (MNC) for unique identification of the mobile network within the country, and the Mobile Subscriber Identity Number (MSIN) identifying the subscriber within the mobile network (a maximum of 10 digits);
- Mobile Subscriber Integrated Services Digital Network (MSISDN): Limited to 15 digits, this number is essentially the mobile phone number assigned to the subscriber. When calling a mobile phone this is the number that is dialled; it contains a Country Code (CC) of up to 3 digits, a National Destination Code (NDC) of 2–3 digits, and the Subscriber Number (SN) *i.e.* the subscriber's phone number (a maximum of 10 digits);
- Short Message System (SMS): Commonly called 'text messages', these have become ubiquitous today as a form of rapid messaging between parties. Early SIM cards could only store around 5 SMS messages, whereas modern SIM cards can store many more. Modern handsets often store these on the handset due to its larger memory capacity;

- Last Number Dialled (LND): This table shows recent calls made and received from the phone, including the time of call. On modern handsets this seems to be mainly stored in the handset memory, with only the most recent call records stored on the SIM;

- Contacts: The contact list for the SIM card is a table of the name and corresponding phone numbers of contacts stored on the SIM card. Early SIM cards could store around 20 contacts, but modern SIMs can store many more, often in the hundreds;

- Location Information (LOCI): This contains the location that the phone was last active in (including CC, MNC and specific Location Area Identifier (LAI) network cell code. Each network cell has its own unique identifier code, moving into to a new cell or turning the phone on will generate a location update request and overwrite this information;

With a large amount of data available about the user and their activities, it is no wonder that forensic investigators seek to examine smart cards. For SIM cards the data that can be obtained includes crucial information on the personal/social/business network connections (address book, recent call register, SMS list) of a user. Other smart cards, including some countries' identity cards, can include financial transactions, public transport use, identity documents, and many more pieces of useful data. Combined with the ability to pinpoint a user's activity to a specific time and location using hidden data within the smart card, as well as corresponding back-end network data, and you have a large assortment of forensically crucial data available for analysis.

Bibliography

[1] M. Ugon, "Portable data carrier including a microprocessor," publication Number: 4211919; Publication Date: 1980-07-08; Application Number: 05/936,694. [Online]. Available: http://patent.ipexl.com/US/4211919.html

[2] S. Kume, I. Yamada, K. Watari, I. Harada, and K. Mitsuishi, "High-Thermal-Conductivity AlN Filler for Polymer/Ceramics Composites," *Journal of the American Ceramic Society*, vol. 92, no. 24730, pp. S153–S156, Jan. 2009.

[3] S. Murali and N. Srikanth, "Acid Decapsulation of Epoxy Molded IC Packages With Copper Wire Bonds," *IEEE Transactions on Electronics Packaging Manufacturing*, vol. 29, no. 3, pp. 179–183, 2006.

[4] H. Dreifus and J. T. Monk, *Smart Cards – A Guide to Building and Managing Smart Card Applications.* John Wiley & Sons, 1998.

[5] A. Steffen, "Smart cards," February 2009, Lecture notes for 'Internet Security 2' course.

[6] D. Kahng and S. Sze, "B.s.t.j. briefs: A floating-gate and its application to memory devices," *The Bell System Technical Journal*, vol. 46, no. 6, pp. 1288–1295, 1967.

[7] R. Kirisawa, S. Aritome, R. Nakayama, T. Endoh, R. Shirota, and F. Masuoka, "A nand structured cell with a new programming technology for highly reliable 5 v-only flash eeprom," in *1990 Symposium on VLSI Technology. Digest of Technical Papers.* IEEE, June 1990, pp. 129–130.

[8] S. Sze and K. K. Ng, *Physics of Semiconductor Devices*, 3rd ed. Wiley Interscience, 2007.

[9] G. Liechty and E. Hirsch, "Lucky-electron model of channel hot electron emission," in *1979 International Electron Devices Meeting*. IEEE, 1979, pp. 22–25.

[10] S. Tam, K. Ping-Keung, and H. Chenming, "Lucky-electron model of channel hot-electron injection in MOSFET's," *IEEE Transactions on Electron Devices*, vol. 31, no. 9, pp. 1116–1125, 1984.

[11] W. Shockley, "Problems related to p-n junctions in silicon," *Solid-State Electronics*, vol. 2, no. 1, pp. 35–60, January 1961.

[12] A. Modelli and B. Ricco, "Electric field and current dependence of sio2 intrinsic breakdown," in *1984 International Electron Devices Meeting*. IEEE, 1984, p. 148.

[13] E. Harari, "Dielectric breakdown in electrically stressed thin films of thermal sio2," *Journal of Applied Physics*, vol. 49, no. 4, p. 2478, April 1978.

[14] D. DiMaria and D. Kerr, "Interface effects and high conductivity in oxides grown from polycrystalline silicon," *Applied Physics Letters*, vol. 27, no. 9, p. 505, November 1975.

[15] L. Faraone, "Thermal sio2 films on n+polycrystalline silicon: Electrical conduction and breakdown," *IEEE Transactions on Electron Devices*, vol. 33, no. 11, p. 1785, November 1986.

[16] S. Mori, Y. Kaneko, N. Arai, Y. Ohshima, H. Araki, K. Narita, and E. Sakagami, "Reliability study of thin inter-poly dielectrics for non-volatile memory application," in *28th International Reliability Physics Symposium 1990*. IEEE, 1990, p. 132.

[17] R. Shiner, J. Caywood, and B. Euzent, "Data retention in eproms," in *18th Annual Reliability Physics Symposium, 1980*. IEEE, 1980, p. 238.

[18] N. Mielke, "New eprom data-loss mechanisms," in *21st Annual Reliability Physics Symposium, 1983*. IEEE, 1983, p. 106.

[19] D. Ielmini, A. Spinelli, and A. Lacaita, "Recent developments on Flash memory reliability," *Microelectronic Engineering*, vol. 80, pp. 321–328, 2005.

[20] S. Aritome, R. Shirota, G. Hemink, T. Endoh, and F. Masuoka, "Reliability issues of flash memory cells," in *Proceedings of the IEEE*, vol. 81, no. 5. IEEE, May 1993, pp. 776–788.

[21] Y. Haghiri and T. Tarantino, *Smart Card Manufacturing - A Practical Guide*. John Wiley & Sons, 2002.

[22] P. Forstner, *Application Report SLAA334A: MSP430 Flash Memory Characteristics*, Texas Instruments, September 2006.

[23] K. Venkat and U. Haensel, *Application Report SLAA392: Understanding MSP430 Flash Data Retention*, Texas Instruments, March 2008.

[24] A. D. Purporti Jr. and J. McElroy, "Full-scale house fire experiment for interfire vr - report of test fr 4009," U.S. Department of Commerce, NIST, Gaithersburg, MD, Tech. Rep., 1998. [Online]. Available: http://www.interfire.org/features/fire_experiment.asp

[25] H. Luhn, "Computer for verifying numbers," Filing Date: 6th January 1954. Issue Date: 23rd August 1960. [Online]. Available: www.google.com/patents/about?id=Y7leAAAAEBAJ

Chapter 4

Sample preparation

The previous chapter outlined in some detail the physical structure and characteristics of smart cards and their non-volatile EEPROM (and flash variant) memory arrays, and the motivation behind examination in a forensic context. This chapter outlines the non-traditional approach to sample preparation taken in this study.

The first step when examining a SIM card is to inspect it for damage – specifically the area in close proximity to the chip module and contact pads. If the sample appears undamaged, or damage is located far from the microprocessor, then standard electronic methods can be used to interrogate the SIM and retrieve data. If the SIM card appears damaged and/or electronic interrogation at this stage fails, then it will be necessary to begin decapsulation of the chip module from the card body. The cryptographic microprocessor is encased within an epoxy moulding compound, and this protective layer will need to be removed to further examine and process the die.

With the microprocessor extracted and clean of any epoxy residue, a more detailed inspection of damage can take place using an optical microscope. If the die is undamaged, then the sample can be re-mounted onto a prototyping chip carrier, contact wires connected to pads on the die, and electrical interrogation can be attempted again. Should the chip be damaged, or electrical interrogation

fail this time, the sample can be further processed in preparation for examination using an AFM.

The memory array stores the data, consisting of stored charges, within the floating gates at the lowest level of circuitry built upon the silicon wafer. To access the floating gates from above is very difficult due to the heterogeneity of materials and structures that constitute the circuit layers. This creates a multitude of problems in trying to remove individual layers of different materials, all while trying to keep the surrounding interpoly dielectric layer intact. A topside approach is too destructive to the floating gate structures, as it often involves using hydrofluoric acid-based etchants to remove higher silicide layers, a chemical that is effective at etching oxide layers such as the IPD.

An underside approach to expose the floating gate tunnel oxide layer [1], however, must only deal with the removal of a single homogeneous material, silicon. To accomplish this, the extracted chip must be re-encapsulated within a new epoxy resin mount to protect it during bulk silicon removal. Once mounted, the silicon can be removed first mechanically through lapping, then chemically with a selective silicon etch, leaving the thin protective tunnel oxide layer intact and exposed. At this stage, the sample preparation is finished and examination with an AFM can begin. The end result could be described simply as microprocessor circuitry, removed of its bulk silicon backbone, and flipped upside-down for underside examination.

A note on populating samples

The early work carried out into decapsulation methods predominantly used PIN locked and otherwise inaccessible samples: as a result, any data populating the SIM cards' memory was present to begin with. Later samples had been populated using a number of custom sample data sets adapted to fill different sized memory arrays. Samples were populated using a USB SIM reader, and free SIM editor software packages.

4.1 Early SIM card evidence analysis

Early on in this study, a piece of evidence came into our possession, the broken SIM card shown in Fig. 4.1, recovered (albeit heavy-handedly) from a covert listening device planted in a businessman's office in Geneva, Switzerland. The SIM card was clearly in two parts, snapped along a break line which possibly ran through the die. Given the state of the SIM card, the first course of action was to ascertain the damage, whether the die had cracked along with the card body.

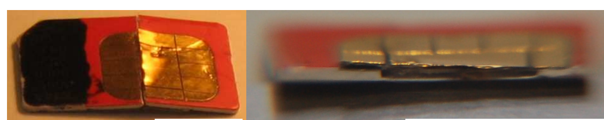


Figure 4.1: Photographs of broken SIM card (left) and magnified view (right) of chip module area (scale bar=1cm).

Initial optical examination of the snapped SIM card revealed that the bond wires connecting the chip to some of the external contact pads had been severed with the snapping of the card body. The wires were visible, still suspended in the resin glob top. The worst case scenario was that the chip itself had been snapped. There also appeared to be a large dent and scrape mark in the centre of the contact pads and the SIM card IMSI (printed on the card body) had been blacked out (unsuccessfully) with permanent marker pen.

Optical microscopy was unable to reveal sufficient information on the state of the die, and so further analysis was conducted using a scanning electron microscope fitted with an energy dispersive X-ray spectroscopy module.

4.1.1 Energy dispersive X-ray spectroscopy

Energy dispersive X-ray spectroscopy (EDS, also EDX) is a non-destructive technique used to characterise a sample's elemental composition. The principle behind the technique is outlined in Fig. 4.2. Normally, an atom contains ground-state electrons in discrete energy shells. Focusing an electron beam onto the

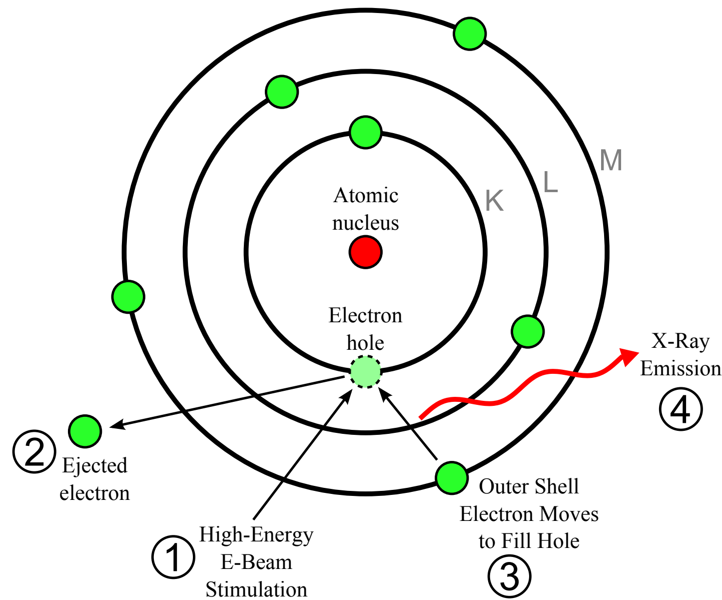


Figure 4.2: Diagram outlining the principle behind energy dispersive X-ray spectroscopy.

sample can excite an electron from an inner shell (K, L, M) of a sample atom, ejecting it and creating an 'electron hole'. An electron from an outer (higher energy) shell moves to fill this hole, with the difference in energy between the two shells being released as an X-ray. These X-rays have characteristic energy levels determined by the energy difference between the two shells, and also specific to the atomic structure of the emitting element – this allows the elemental composition of the sample to be identified by examining the spectra of emitted radiation. These systems are most commonly found in SEM setups due to the overlap in required components.

The EDS X-ray detector, typically a lithium-drifted solid-state silicon device, measures the relative abundance of emitted X-rays and their energies. When an incident X-ray strikes the detector, it generates a charge pulse in proportion to the X-ray energy. The charge pulse is converted to a proportional voltage pulse by a charge-sensitive pre-amplifier and is then sorted by a multichannel analyzer and sent to a computer. The X-ray energy spectrum and its respective count-rates can then be evaluated to determine the elemental composition.

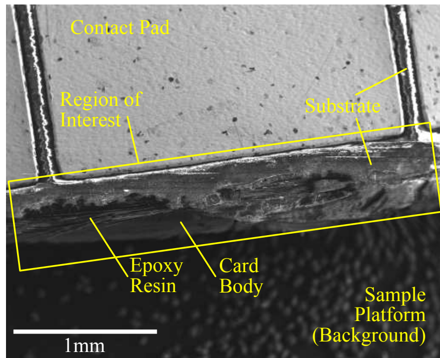
It is possible to conduct EDS in a variety of modes depending on X-ray collection and the area 'illuminated' by the incident electron beam. Qualitative analysis compares the spectral peaks to known X-ray energy values to determine the presence of an element. Elements from beryllium through uranium are capable of being detected with minimum limits varying from 0.1 to a few percent composition depending on the sample structure and element in question. More quantitative analysis is possible from the relative X-ray counts and with calibrated samples of similar structure, a much greater accuracy can be obtained over that of an unknown structure.

It is possible to scan the electron beam along a line and detect X-ray emission at discrete positions, thus forming a linear profile of a sample's composition. This is particularly useful for the analysis of electronics and integrated circuit analysis. Alternatively, an entire area can be mapped in the same way, allowing for individual elemental composition maps showing local relative concentration through colour/brightness intensity. A lateral resolution of around 1μm is possible in this mode.

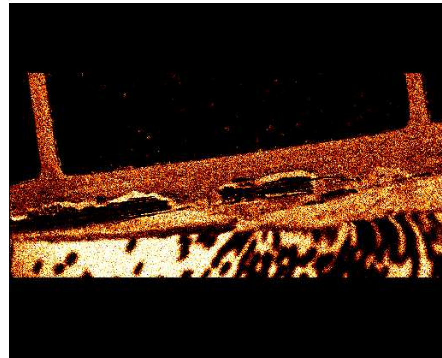
4.1.2 Exhibit examination

To ascertain whether the silicon chip had been broken, EDS was conducted. Assuming that the chip was at the centre of the chip module area, underneath the central contact pad like other cards, SEM examination of this half of the snapped SIM would be impossible – charge-induced damage from the electron beam to circuitry and/or stored data held within would be unacceptable. However, the decision was made to inspect the other half of the snapped SIM (the half with the blacked out IMSI). The initial SEM micrograph taken over an area at an angle showing the cross section of this half of the snapped SIM's chip module area is given in Fig. 4.3a with parts labelled.

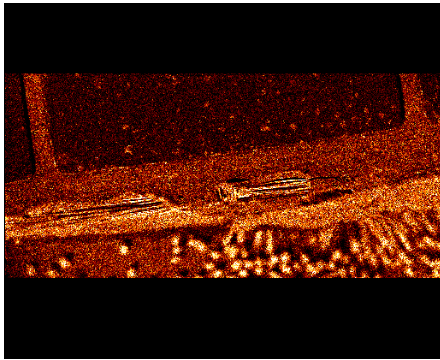
Figure 4.3 shows X-ray maps of relative concentrations of elements in various areas of the SIM card chip module area. The contact pads, as expected, have



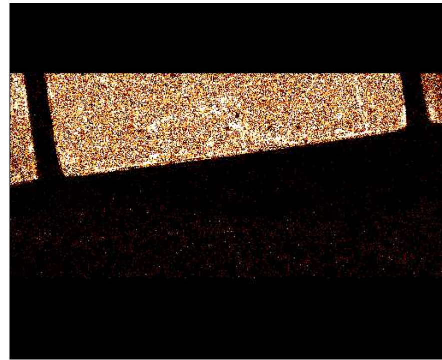
(a) SEM micrograph



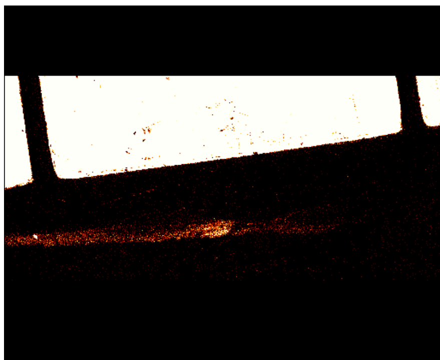
(b) Carbon



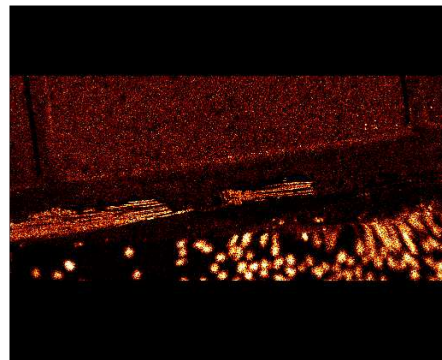
(c) Oxygen



(d) Nickel



(e) Gold



(f) Silicon

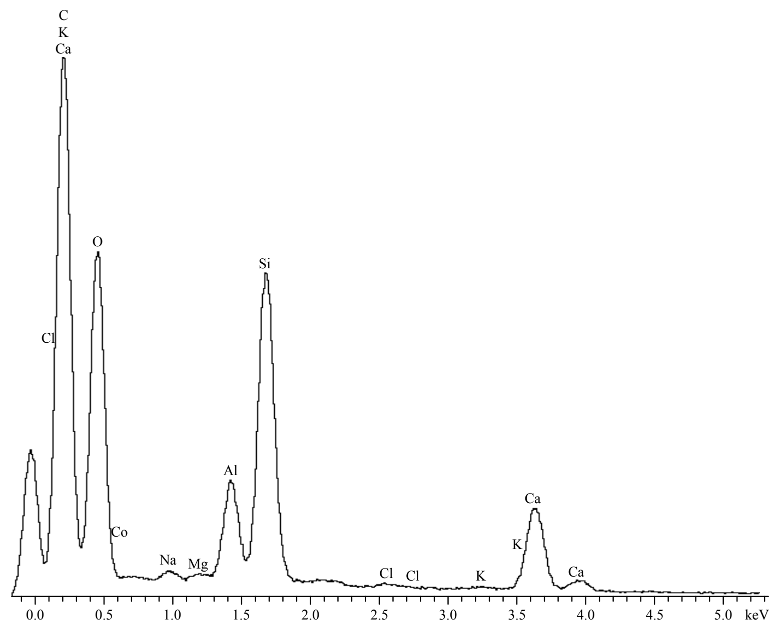
Figure 4.3: SEM image and corresponding EDS X-ray maps showing relative concentrations of elements in various areas of the broken SIM card.

high relative concentrations of gold and nickel, see Figs. 4.3d and 4.3e. Beneath and between the contact pads is the substrate layer with high concentrations of carbon, Fig. 4.3b, and oxygen, Fig. 4.3c. The 'stranded' area visible beneath the substrate layer could either be chip adhesive around the edges of the die or more likely: the epoxy encapsulant 'glob-top'. It has a structure consisting of high levels of oxygen and silicon with non-uniform concentrations, Fig. 4.3f. Further quantitative EDS analysis, Fig. 4.4a, reveal peaks in the spectra identifying the presence of carbon, oxygen, aluminium, and silicon – the vast majority of elements found in an epoxy resin filler material, see Table 3.2. The levels of silicon are substantially lower than would be expected should this be a broken-off piece of the chip.

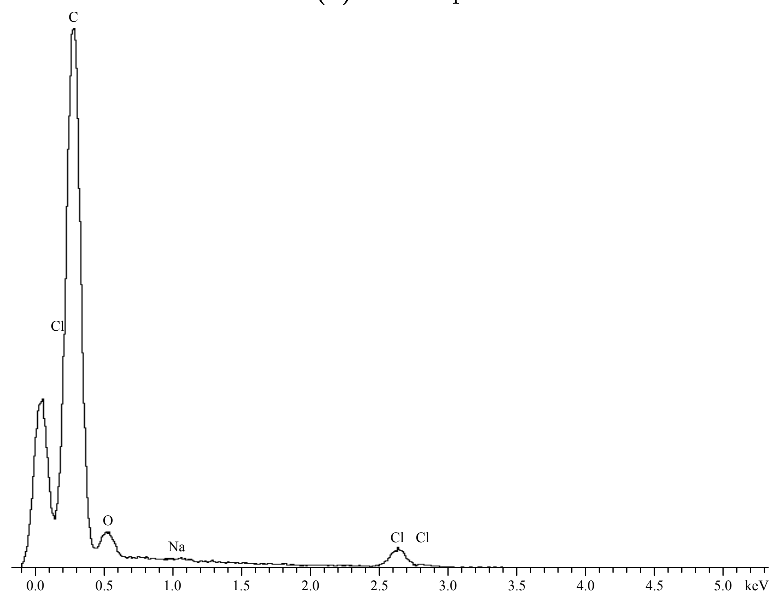
The lowest visible structure (above the sample platform background) has high levels of carbon indicating it is the card body. The spectra for this region, Fig. 4.4b, shows that this is likely to be PVC; but could possibly be polycarbonate, with the low secondary chlorine peak explained by phosgene remnants within the bulk plastic from manufacturing – in either case confirming that this is the card body being analysed.

Since there was no obvious silicon presence where one would expect to find it should the chip have been snapped, it was decided that the chip was likely to be all in the other half. The chip would still need to be removed, inspected for damage and have new bond wires connected to allow electronic interrogation to take place.

Further mechanical decapsulation revealed that the SIM die, despite not having been snapped in half along with the card body/chip module, had indeed suffered extensive damage (Fig. 4.5a). This appears to have been caused in conjunction with the dent and scratch visible in the top central contact pad (directly above the chip). The fracture seems consistent with damage caused by a screwdriver or punch-like tool in a single slightly-angular blow to the central pad; Figs. 4.5b and 4.5c shows the 'scar' fractured through to the underside of

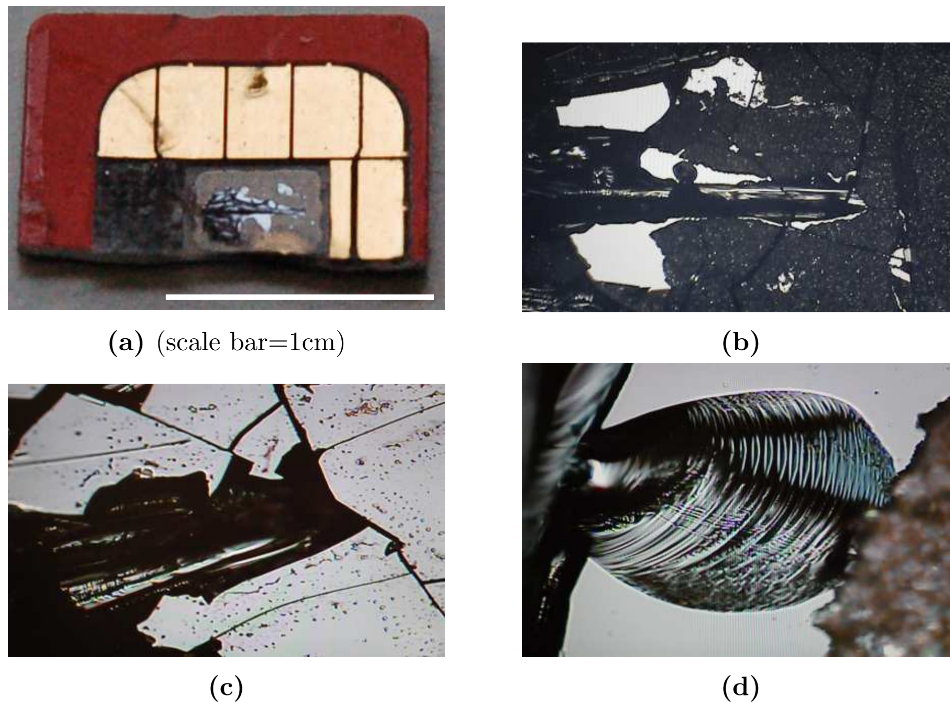


(a) Glob top



(b) Card body

Figure 4.4: EDS spectroscopic analysis spectra for glob top and card body sections.



(a) (scale bar=1cm) **(b)**

(c) **(d)**

Figure 4.5: Photograph of the SIM card silicon damage after central contact pad removal **(a)**, with central break **(b)**, extensive silicon damage **(c)**, and conchoidal (clamshell) break around striking point.

the silicon chip. Also visible are conchoidal shock fractures common in cases of quartz and silicon damage, see Fig. 4.5d.

Finally, full removal of the die from the epoxy glob top using DePOT (see Section 4.2) revealed the full extent of the damage – the silicon was being held together by the epoxy glob top and upon dissolution the depth of the fractures was revealed, the die literally fell to pieces. As a result of this extensive damage to the silicon die, electronic retrieval of any information was impossible from this piece of evidence.

4.2 Resins: decapsulation

Epoxy resins are thermosetting polymers containing more than one epoxide group. They are used as adhesives, high-performance coatings and encapsulating materials in many industries, *e.g.* aerospace engineering, electronic/electrical

component manufacture, industrial tooling, dentistry. To set the epoxy into a rigid plastic material, a curing agent (hardener) is mixed to an optimal ratio with the epoxy. A reaction occurs between the hardener and the epoxide groups which generates heat and hardens the mixture to form a highly cross-linked 3D network. Epoxy resins cure at a range of temperatures from 5–150°C, depending on the specific epoxy curing agent, but in general, those epoxy resins cured with heat will be more heat and chemical resistant than those cured at room temperature.

The semiconductor industry uses epoxy resins in the majority of cases to protect delicate electronic chips due to their outstanding chemical resistance, high physical durability and low porosity – precisely the properties that those practising failure analysis need to overcome. To overcome the brittle nature of epoxies, various additives are included in the mixture; a list of the typical constituents of epoxy resins is mentioned in Table 3.2 in Section 3.2. Inclusion of these additives to improve upon any negative properties of the epoxy, *e.g.* brittleness, tends to reduce the effectiveness of existing positive properties, such as chemical resistance.

A variety of methods exist for removing the silicon chip from the protective encapsulation depending on the weaknesses found in the packaging materials, methods discussed in literature range from direct chemical extraction [2] to heating to weaken the chip module adhesive [3]. Unfortunately, this is not a simple task; the encapsulants are designed to be highly resistant to both physical and chemical attacks. Given that this is usually a precursor in the field of semiconductor failure analysis, few papers focus solely on decapsulation, let alone give extensive detail on the process involved.

The removal of an epoxy typically occurs in two stages, the first being substantially slower than the second. To begin with, only the outer layers of the epoxy are attacked, the speed of which depends upon the extent of 'blushing' the resin underwent during manufacture. 'Blush' is a wax-like coating that forms

during the curing process due to oxygen and moisture in the atmosphere. This blush layer has fewer exposed OH groups than the underlying bulk epoxy and hence forms a far tougher surface layer for acids or solvents to react with. This is where the second stage of the removal occurs; typically much faster and care must be taken to avoid damaging the encapsulated chip beneath.

An established technique includes the use of hot acids or acid blends to dissolve dual in-line (DIP or DIL) packaging. Early guides to these acid attacks were given by W. Joe Byrne of the National Semiconductor Corporation [4] and M. Jacques of Teledyne Electronics [5]. It is very common to use hot concentrated sulphuric acid (H_2SO_4) for its low cost and reliability in exposing the die surface [6]. In Byrne's paper it is stated that the success of the technique relies on the use of high temperature (220–250°C). The thorough dehydration of the heated acid is key to the selective removal of epoxy whilst leaving aluminium metal intact. However, as the epoxy decomposes, water is released and aluminium will be corroded. Concentrated sulphuric acid is specifically suited for the removal of silicone moulding compounds.

To initiate the technique, a small hole should be carefully drilled into the packaging above the die and the part should be submerged into the hot acid bath for 20–30 seconds. A thorough cleaning process consists as follows: immersion into a room temperature acid bath to slow the etch process, followed by large amounts of DI water, followed by acetone and finally blown dry with an air or nitrogen gun. This process is vital as any acid remaining on the surface can dissolve in moisture from the atmosphere and cause further uncontrolled decomposition.

The second decapsulation technique involving acid outlined in Byrne's paper uses 90% fuming nitric acid (HNO_3) heated to 70°C and follows much the same process except that only acetone is used in cleaning – immersion in water to quench the acid etch will increase the reactivity of nitric acid with any exposed aluminium metal on the surface. Kuhn's paper [3] also uses fuming nitric acid

and suggests that at 60°C it is suitable as it attacks epoxy whilst leaving gold, aluminium, SiN, Si, and SiO_2 intact. The recommended subsequent wash is first to use acetone, followed by DI water, then IPA, and finally blown dry.

In discussing this technique on larger samples, Kuhn's paper references 'Integrated Circuit Failure Analysis: A Guide to Preparation Techniques' [7], a book by Friedrich Beck which covers this area extensively. Beck states that *"the number one commandment in wet chemical opening of plastic packages is to keep moisture away"*. A well should be milled and the sample should be heated to approximately 60°C. Drops of fuming nitric acid at around 60–70°C should be dropped into the milled well in the heated sample and left to work for 10 seconds, then rinsed with a spray of acetone to remove reaction products and exposed filler. This process is repeated until the well is large enough to access the die.

Jacques recommends the use of glacial (pure, water-free) acetic acid as the first rinse step each time to remove packaging material. This is followed by a rinse in acetone to remove the excess acetic acid and placed back onto a hotplate to evaporate the acetone and inspect the sample. Beck, in contrast, uses only acetone, but recommends the use of an extensive clean afterwards: ultrasonic clean in acetone for 20 seconds, then 20 more seconds of ultrasonic cleaning in DI water to remove salt residues, and a further ultrasonic clean in IPA to expel water.

De Nardi *et al.* [8] mention a higher temperature process to remove packaging around EEPROM memory chips carried out at 90°C and refer to their colleague Perdu's earlier paper [9]. They mention immersion into 90°C fuming nitric acid for 4–5 mins. Perdu's paper offers an insight into a multitude of different decapsulation techniques. Many techniques are mentioned as well as equipment names and models. Unfortunately, there is little in the way of detail in the paper, especially with regards to their performance in different applications. Wet etching was carried out by Perdu with a jet etcher, an automated piece of equipment that heats the fuming nitric acid and automatically applies it in precise bursts

to a sample. Murali [10] covers the use of jet etchers in far more detail, outlining various parameters and commenting on their performance. Parameters such as flow rate, temperature, volume, and acid mix composition are all investigated and observations are made with respect to completion of decapsulation and also the propensity to attack copper bond wires.

A publication by Jiang *et al.* [11] on die cracking during destructive analysis is worth noting. Tests were carried out to ascertain whether acid soaking caused die cracking in multiple chip packages and the results were surprising; after soaking, even for long periods in near-boiling nitric acid (>86%) for up to 2 hours (commonly their dies only require 20–30 mins of etching) the dies remained intact, but at lower etch temperatures the expansion and redistribution of stress throughout the silicon could account for an almost 100% process failure rate.

Other methods exist beyond those mentioned already; Perdu's paper briefly mentions some of these techniques. Reactive ion etching (RIE) is a viable technique, albeit not in cases where data integrity is paramount due to the charging effects of ion bombardment interfering with the floating gate potentials and disrupting stored data. RIE is far slower than wet chemical decapsulation techniques, with an etch rate of 1.5–3μm/min for standard epoxide moulding materials. This process can take several hours to etch away around 1mm of material (a common stopping point if initially milling an entry cavity in the package.

Laser ablation (otherwise known as photoablation) is another relatively new technique in failure analysis. Like RIE, it is far slower than wet chemical etching. The first reported use, by Guo *et al.* in 1995, was in the ablation of a block of pure graphite [12] and shortly thereafter graphite mixed with a catalytic metal [13]. Used for decapsulation, the technique offers a reliable and accurate approach to die removal, however, thermal stresses are induced in the die. It should be noted that because of the thermal damage caused by the laser, in particular the melting of the metallisation [14], this method is most often used as a precursor to traditional wet chemical etching, reducing the time needed for an acid soak. By

removing packaging to create a uniform thickness above the die the etch front during the shorter acid soak reaches the die surface more uniformly, avoiding any over-etching [15].

4.2.1 Thermal/mechanical decapsulation

Safe and effective extraction of SIM card microprocessors is the first stage along the path to examining their EEPROM memory arrays. Experiments were conducted on numerous samples of different manufacturers, types and different encapsulation materials to develop a reliable and forensically sound method of die extraction. The first attempts made to decapsulate SIM cards were purely mechanical; craft knives were used to trim the excess card material away, exposing the chip module. Carefully cutting along the edge of the die, the aim was to expose one edge side and from there work to carefully peel off the glob top. Given the physical resistance of the epoxy encapsulant material, simply cutting away the glob top with craft knives failed to successfully remove any of the early SIMs intact.

The next method tested to remove the epoxy encapsulant came from this project's initial feasibility study, carried out by Jones and Kenyon [16]. Jones stated that in some cases it proved possible to decapsulate the chip module mechanically in combination with gradual 'heating to temperatures not exceeding 160°C', whereby the glob top became more pliable.

Heating a sample set of 14 assorted SIM cards, consisting of 8 different varieties from different manufacturers, to temperatures in accordance with Jones' findings found that this method had a success rate ≈20%. Three samples appeared to remain completely unchanged, and 4 samples actually appeared to harden, turning a dark brown colour. In only 7 cases did the epoxy glob top become softened from heating, and only in 3 of these was decapsulation relatively successful (albeit retaining some epoxy on the die surface). In 4 of the softened epoxy cases the glob top was not able to be fully removed and instead snapped

Table 4.1: Depot 1 constituents.

Material	Percentage (by weight)
Dimethyl sulfoxide	70–90
1-phenoxy-2-propanol	10–30

along the edges of the die, leaving the epoxy directly above the die intact and with no vantage point to effect its removal.

In spite of these results, an extremely effective technique for the initial separation of the PVC card body from the chip module was discovered. Placing the whole SIM card, contacts-side down, onto a hotplate heated to between 120–150°C for 3–5 seconds softened the hot melt adhesive between the card body and the chip module very quickly. After this heating, the two parts could be carefully pulled apart using two pairs of tweezers.

4.2.2 Solvent decapsulation

Specifically concerning the decapsulation of smart cards, existing forensic techniques include the use of solvents. There are also specially-formulated blends of solvents available to remove the epoxy resin – 'DePOT' is one brand of such blends. DePOT are reusable, pH neutral blends of solvents from Crownhill Associates Ltd. [17]. They are stated to be able to disintegrate a range of common epoxies such as Fiberite E9451, Pacific Resins EMC90 and Hysol MG5F, plus numerous epoxy casting compounds. Two blends of DePOT were available at the time of testing, and a third water-based DePOT 3 blend is available now. Each blend is suited to attacking different types of epoxy, but it worth noting that all types of DePOT will also attack most thermoplastics and a wide range of thermosetting materials. DePOT 1, Table 4.1, is typically used hot, to a maximum of 150°C, DePOT 2, Table 4.2, is used at lower temperatures, typically around 30–40°C.

Various individual solvents were tested for their ability to soften/dissolve

Table 4.2: Depot 2 constituents.

Material	Percentage (by weight)
Benzyl alcohol	30–40
Dipropylene glycol	5–10
Light aromatic naptha depleted solvent	5–10
Hydroxyacetic acid	1–5

Table 4.3: Decapsulation using solvents: week-long room temperature soak test observations.

Solvent	Observations
DI water	No effect
Ethanol	No effect
Acetone	Softened most samples
Isopropyl alcohol (IPA)	No effect
Methyl isobutyl ketone (MIBK)	EMC softened slightly
Toluene	Slight softening for a couple of samples
Methyl ethyl ketone (MEK)	Softened most samples
DePOT 1	No effect
DePOT 2	EMC slightly softened for a few samples

EMCs at room temperature during a prolonged week-long soak. The most effective of the solvents were acetone and methyl ethyl ketone (butanone), although results varied depending on the sample tested, observational results are shown in Table 4.3.

Acetone is a good solvent for most plastics and synthetic fibres, is commonly used as part of cleaning agents such as nail polish remover and superglue remover, and can also be used for dissolving two-part epoxies before hardening. Soaking in acetone resulted in a range of effects observed. Some samples became moderately softened, others showed no effect, and in a few the glob top became noticeably swollen. In these cases the epoxy glob top softened and acetone seeped in behind the epoxy, against the surface of the die. This resulting bubble over the die circuitry separated the EMC from the die surface, making acetone a rather effective epoxy remover in those cases. Methyl ethyl ketone (MEK) exhibited a similar, but less notable effect.

Further tests with acetone, showed that this swelling does not always extend

Figure 4.6: Photograph of the circuit-side of a decapsulated SIM chip still retaining a large amount of EMC on surface of topside circuit.

to cover the whole surface of the die, nor does it always begin at circuit-epoxy interface. As was discovered after further mechanical decapsulation, sometimes the bubble formed mid-way through the resin layer, only resulting in the easy mechanical removal of part of the epoxy, and leaving behind a thin layer on the die surface. Blends of acetone with other solvents (isopropanol, butanone, MIBK) were in all cases mildly less effective than pure.

The relative ineffectiveness (taking a week to soften an epoxy glob top) using solvents stemmed from the lack of aggressiveness in the process. The end result, if successful, is a partially clean sample – some EMC is often left on the top surface of the die, see Fig. 4.6, and can cause major problems further along the sample preparation process. Those samples with a noticeably 'blushed' epoxy layer were mostly unaffected by these solvent attacks.

To improve the effectiveness of the solvents, agitation in an ultrasonic bath was performed for 1 hour per sample in their respective solvents. The solvents were topped up as needed throughout the agitation period to replace any lost through effervescence and evaporation. The aim of this was to increase the dissolution rate of the solvent and aid in the removal of by-products from the reaction surface. In all solvents that previously showed some ability to dissolve epoxies, the ultrasonic agitation (and its subsequent heating effect) improved their effectiveness. Of particular note was DePOT 2, designed to be used at 30–40°C; this blend exhibited a much greater effectiveness when used in concert

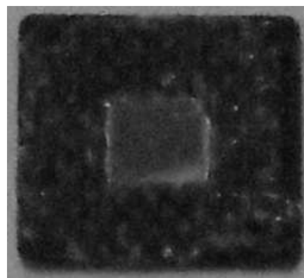


Figure 4.7: Photograph of a SIM card chip module, removed from card body, of a type unable to be dissolved in solvents or solvent blends. The inner square is the exposed underside of the die; the outer structure is the chemically-resistant encapsulant.

with ultrasonic agitation.

The most effective solvents were acetone, MEK and DePOT 2, but none of these managed to fully dissolve the last remnants of EMC on the surface of the chips – optical inspection revealed that in all cases some epoxy remained on the surface of the die. Repeating the agitated soak a second time did little to improve on the surface cleanliness. SIM cards with chip modules such as that pictured in Fig. 4.7 remained unaffected by any solvents or blends.

At the Forensic Science Service (FSS), SIM cards that were damaged beyond electrical interrogation would be decapsulated using DePOT 1 (see Table 4.1) by forensic examiners. The solvent blend was placed in a beaker and heated with a hotplate set to approximately 120°C. The aim was for the solvent blend to dissolve the card body and epoxy glob top, leaving the chip module area otherwise intact – including bond wires. The chip module was carefully washed in water and optically inspected for damage. Undamaged chip dies could be reconnected where necessary, *i.e.* manually reconnecting a bond wire that had detached from the large contact pad using silver paste and tweezers. However, if the SIM die itself was damaged or the bond wires were severed too close to the die (making the bond wire too short for rewiring to the external pad), they ceased processing and judged it unable to be electrically interrogated.

Understandably, this primitive and underdeveloped method had a very low

success rate and relied far too heavily on random chance, that if a bond wire broke too close to the die during decapsulation, the evidence stored would be lost forever. From discussions with examiners in the Electronic Forensic Sciences Unit (EFSU), they knew this was not the case, and had read into possible techniques for improving and developing this method, but they were confined by time, budget, resources and 'executive decisions' to continue with the status quo – after all, less than '1-in-100' exhibits required this level of processing; most came in unscathed, with handsets or alone.

Tests conducted with DePOT 1 at 130–140°C showed it to be the most effective decapsulation method tested thus far. After initial tests as to its efficacy, Depot 1 was able to rapidly and effectively dissolve almost 80% of the EMCs encountered in SIM cards, including some of those unaffected, or only partially dissolved by previous solvent tests. However, inspection under an optical microscope revealed that even these 'successful' samples were not fully clean. Furthermore, this blend of solvents had no effect on certain samples, such as that shown in Fig. 4.7 – now thought to be an epoxy that is specially formulated to have an increased chemical resistance. Other samples which took longer to dissolve were those with a higher level of blushing – it became clear that a more aggressive process was required to obtain access to such samples, and to obtain cleaner dies overall.

4.2.3 Fuming nitric acid decapsulation

An investigation was carried out with the aid of Dr. Andrea Sella[1] into the use of fuming nitric acid (≥99%) as a potential decapsulation process. A small (5ml) beaker of fuming nitric acid was heated on a hotplate to an estimated 90–100°C (boiling point of 120.5°C). Samples were pre-prepared only to the point of separating the chip module from the card body using the previously mentioned hotplate method (a method which can take place along-side this process given

[1]Department of Chemistry, University College London

the hotplate used). The chip modules were held in the acid with tweezers, care was taken to avoid applying pressure to the die and instead the substrate and contact pads were gripped. The samples were inspected for progress every 10 seconds.

Most samples behaved as follows: after around 15s the blush layer was breached and more vigorous reactions were observed; after 30s most of the surrounding encapsulant was removed, a thin layer of encapsulant is still visible; after 60s the chips were mostly clean of epoxy, but close examination under a microscope revealed a layer still present on the chip topside; after 90s decapsulation was complete, though there remains a layer of etch residue left across the entire chip module.

Various methods at washing away the residue left on the surface of the die were tested. Both DI water and acetone were found to be ineffective as first rinses to remove the fuming nitric acid residue. Fuming nitric acid has little or no effect on the exposed aluminium contact pads and gold bond wires due to low water content, but when DI water was used at a first rinse, Al etching reactions were observed by optical microscopy – indicating that the water was not only insufficient at washing away the nitric acid residue, but that it acted to dilute it, thus initiating corrosion of the metal. Acetone was found to be an ineffective first rinse without the aid of ultrasonic agitation.

Glacial acetic acid, as mentioned by Jacques [5], was found to be a superb initial rinse, requiring only a small amount – between 3 and 5 drops carefully dripped across the sample surface will suffice. After this initial glacial acetic acid rinse, a series of soaks at room temperature were used: acetone, DI water, and IPA soaks for 2 minutes each. The sample should then be gently blown with either a compressed air or nitrogen gun to dry the surface.

Those samples encased within an epoxy with a noticeable 'blush' usually took 10–15s longer to decapsulate. In a few cases this took up to 45s longer to overcome this first stage and exhibit the more visible, vigorous reaction of

bulk encapsulant dissolution. As mentioned previously, this is caused by the waxy, chemical-resistant properties of the 'blush' layer and its lack of exposed OH groups. Once this layer was breached, decapsulation continued at a similar pace to other samples.

In addition, those chips that were previously inaccessible, *e.g.* similar to that shown in Fig. 4.7, were successfully decapsulated using this method. Often, these samples were faster to decapsulate using fuming nitric acid than those with more common epoxy encapsulants – taking approximately 30s in total to complete the process. For some EMCs, the time to completely decapsulate was as long as 3 minutes, double the usual expected time; however, no SIM chip module has yet been encountered which cannot be decapsulated in this manner.

The findings are conclusive: near-boiling fuming nitric acid provides a rapid decapsulation technique that works on all types of EMC tested thus far, removing the epoxy but leaving the die and quite often bond wires and external contact pads intact. The chips extracted using this method are much cleaner than those decapsulated with heat, solvents or solvent blends. The use of a small amount glacial acetic acid was extremely effective at removing the remaining acidic residue on the surface.

As with all decapsulation techniques, once completed, a thorough optical inspection of the exhibit is required to assess any damage. If the die appears undamaged, then mounting it carefully onto a prototyping sample holder and using a wire bonder to connecting wires to the contact pads may allow for electrical interrogation of the data. If the chip is damaged, or electrical interrogation fails, then the next step is to mount the chip into epoxy resin for further processing.

4.3 Resins: mounting

At this stage it is necessary to re-encapsulate the microprocessors within epoxy resin blocks, leaving only their silicon undersides exposed. The ideal epoxy

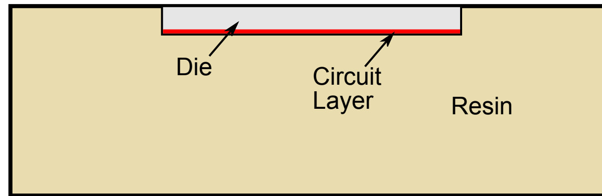


Figure 4.8: Schematic showing the ideal situation of die setting in resin – the die sits level with the base of the cast and no bubbles reside within the resin.

casting is shown in Fig. 4.8 and has the microprocessor sitting parallel to the base of the cast, allowing for lapping parallel to the topside circuitry layer to take place.

The purpose of encapsulation of the naked die after packaging removal is threefold: to allow the small, delicate die to be handled and controlled more easily when set within a larger resin block during lapping/polishing; to expose only the bulk silicon underside for silicon removal, protecting the surface circuitry from being attacked by mechanical or chemical attacks; and to provide a 'backbone' for the remaining fragile die circuit after the bulk silicon etching is complete. A digital thickness gauge was used to measure the thickness of each die prior to mounting, enabling the remaining thickness of bulk silicon after each lapping step to be estimated.

Little noteworthy information is published nowadays on the topic of resin mounting (otherwise known as encapsulation) of devices, especially with regard to protecting samples for failure analysis. There are, however, two noteworthy publications worth discussing. Firstly, Wong's 1995 paper [18] – more a study of common practice and methodology than any notable developments in this field, it was nonetheless a useful starting point since it outlined the various possibilities for encapsulation techniques in industry, including information on 'glob-top' coatings commonly found in smart cards.

Secondly, Liebert analyses possible epoxies for use in failure analysis in some detail in his paper [19]. With much the same requirements as this investigation

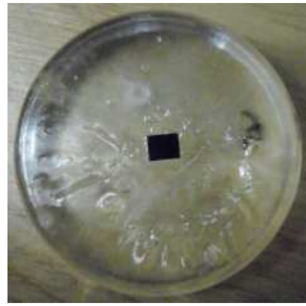
has, this paper has proven very useful in identifying common problems with the encapsulation process. He points out that the epoxy chosen must have a 'similar hardness' to silicon to better protect the die from mechanical damage during parallel lapping. Although this is unrealistic if taken literally, it is clear that if the epoxy is much softer than the sample then it would lead to a rounding of the die edges.

The paper reveals observations that cold-setting, highly-viscous epoxies often formed bubbles and voids around the edges of the die. This was posited to cause problems later on during processing as areas of the die edges are left partially exposed to chemical etchants; larger bubbles could even cause under-etching by exposing parts of the top circuitry. It is possible to limit the aeration through slower, more careful mixing of the epoxy, and of using a thin wire loop to remove larger bubbles before pouring into moulds.

During heating in a silicon etching solution it was found that some cold-setting epoxies tend to expand to a different extent from the encapsulated die, exposing the edges of the die to chemical attack – this is due to older recipe epoxies lacking the mineral filler of the newer recipes. This filler, *i.e.* the spherical silica particles, acts to reduce the coefficient of thermal expansion and are much preferred for use in failure analysis.

Of the hot-setting epoxies (those that are heated and moulded into shape in a moulding press), those in pellet form are also notably porous, offering insufficient chemical resistance and a powdered form is required to adequately protect the die from attack. Furthermore, intense pressure from the moulding press is not properly dissipated across the die when the resin begins in pellet form, and this can induce unacceptable die cracking.

Initial trials into the use of hot-set epoxy resins for the preparation of SIM card microprocessors concurred with these findings. Ten samples were set into an epoxy using hot-set resin granules and a compression mould at the Experimental Techniques Centre at Brunel University. This process resulted in a 50% failure



(a) Resin test with 5:1 mix ratio

(b) Resin test with 5:2 mix ratio

Figure 4.9: Photographs of end result of resin ratio tests.

rate – it became clear that the compression of resin precursor granules was causing excessive and uneven stress on the die during the moulding process. The end result was that half of the samples were shattered during encapsulation.

The resin performed adequately during lapping and polishing, providing a support while adequately protecting the sides of the die from edge-rounding. However, this resin proved to be completely unsuitable for the later stages of processing – wet chemical etching. While the resin appeared sufficiently resistant to the short buffered oxide etch, low chemical resistance to heated hydroxide solutions led to a rapid, and rather spectacular disintegration of the resin. The resin rapidly dissolved into solution, forming a viscous and completely opaque etching solution within seconds. The thinned dies were released from the resin 'pucks' as the resin dissolved, exposing the circuits to the etching solution and thus destroying the samples.

The second epoxy used was Buehler Epo-Thin (2-part) cold-set resin moulded into silicone rubber moulds. This resin offered excellent chemical resistance to both short etches in hydrofluoric acid and prolonged etches in heated hydroxide solutions, as well as low shrinkage. The downsides to this resin lie in the fact that it is a cold-set resin, requiring two parts to be mixed.

The mix ratio of 5:1 (resin:hardener) stated in the instructions for this epoxy was found to be incorrect. Various mix ratios near to that stated in the instructions were tested; leaving them to set at room temperature, the epoxies were

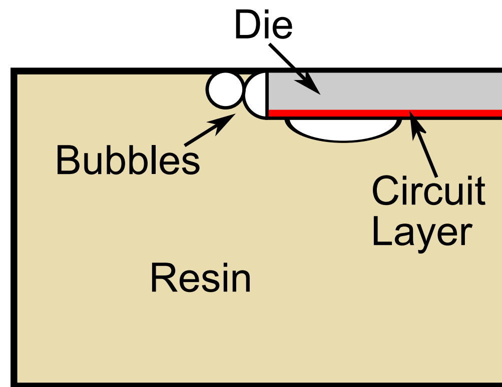


Figure 4.10: Schematic showing common problem with bubble formation in resin casts.

checked periodically. The results were clearly visible, see Fig. 4.9. Those pucks formed from a mix ratio of 5:1 and below has insufficient hardener to set the epoxy in a suitable time, taking at most 10 days to set to a moderate hardness (and in some cases still remaining viscous on the bottom of the silicone mould). With a 5:2 ratio mix the resin had set solid in approximately 10 hours – the expected time for the resin.

The next obstacle to overcome was the formation of bubbles within the resin. While not an issue when in suspension, they had a tendency to congregate on the top surface and around the die edges, see Fig. 4.10. Bubbles on the edges of the die weaken the protection offered by the resin, allowing under-etching during oxide and silicon chemical etches, potentially destroying circuitry and putting stored data at risk. Those on the surface of the die reduce the contact surface between resin and exposed circuitry, leading to a greater degree of delamination problems at the AFM examination stage.

Even with care, stirring together the epoxy and hardener with a stirring rod introduced too many bubbles and was often insufficient at mixing the two parts effectively. A magnetic stirrer was found to mix the resin more effectively, and periodic agitation bursts in an ultrasonic bath were found to aid in lifting the bubbles to the surface of the mix.

Figure 4.11: Photograph of a superior resin cast created through mixing with a magnetic stirrer and agitating for short bursts in an ultrasonic bath.

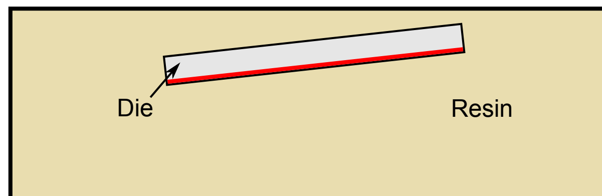


Figure 4.12: Schematic showing problem of die setting at an angle to the mould base.

After a few minutes of magnetic stirring and periodic ultrasonic treatments the mixture was left for a few minutes to stand. This allowed the last of the bubbles to rise to the surface, where the bubbles tend to combine and burst. The remaining bubbles could then be scooped off of the top surface easily prior to pouring into the moulds. Overall, this produced a set of superior resin casts with notably fewer bubbles, as shown in Fig. 4.11.

One final problem that should be discussed with resin encapsulation was when the die rises slightly from the base of the mould, finally being set in resin at an angle to the base of the cast, see Fig. 4.12. Uneven die setting leads to non-uniform mechanical and chemical etching, and the potential destruction of data.

The main culprits behind this are bubbles congregating around the die edges – they act to seep underneath the die, lifting it to an angle that will eventually set and become permanent. By reducing bubbles in the epoxy resin mixture using previous methods, the likelihood of this issue occurring was much reduced.

The use of a 'silicone release spray', designed to aid in epoxy cast removal, lightly sprayed and wiped around the mould interior before mounting appeared to also help somewhat – posited to provide some helpful surface tension between the die and the mould.

When the die is set unevenly it can be difficult to rectify during mechanical silicon removal. Correcting for this angle requires extreme care, attention, and patience during mechanical silicon removal, but it is possible to do so. In cases where this is spotted immediately after pouring, two small, soft plastic rods were successfully used to manipulate opposite corners of the die, rearranging it on the base of the mould.

One possible improvement upon the final process outlined in this study is to use a vacuum chamber. Holding the poured resin inside the vacuum for several minutes will act to draw air out of both internal sample cavities and the resin suspension. This external pressure should collapse the resin into any internal cavities on the die circuit, resulting in a superior fitting mould. The ideal outcome of the entire preparation process is shown in Fig. 4.13a. This outcome is the result of the successful removal of the bulk silicon by wet chemical etching, leaving the topside circuitry sitting flat in the resin puck, exposed and ready for underside examination.

The critical nature of the decapsulation process was discovered when early samples, which had not been fully decapsulated, and thus retained a thin layer of the original EMC, were processed. After bulk silicon removal, the circuitry had a tendency to peel away from the resin, rendering any SPM analysis impossible. This problem is referred to as 'delamination' in this study, and is illustrated in Fig. 4.13b.

One potential solution to this is to deposit another material on top of the cleanly decapsulated circuitry prior to setting into resin; thus providing a structural support for the thin circuitry once the bulk silicon has been removed, see Fig. 4.13c. The main concern raised with this solution lies in the choice of de-

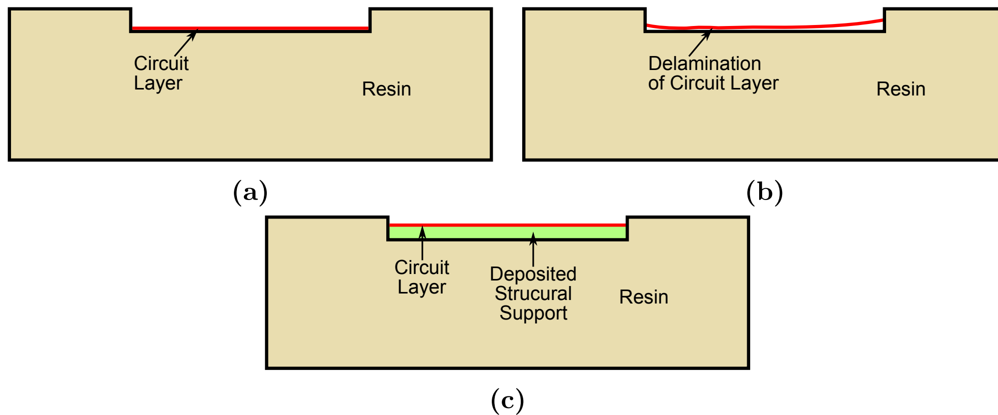


Figure 4.13: Diagrams showing the final outcome of processing. The ideal outcome after wet chemical etching – exposed circuitry set properly into the resin **(a)**, problems with delamination of the circuitry from the resin **(b)**, and a possible solution to delamination – deposition of a new structural support prior to setting in resin **(c)**.

position method to create a new topside support on the circuit. It is assumed that the use of plasma-type methods would be detrimental to the data integrity, however, this needs to be tested. Evaporation deposition methods introduce high temperatures, and these cause issues with data retention, but again require testing before exclusion.

4.4 Mechanical silicon removal

With the die now encased within a new resin 'puck' for protection, the next step is to remove most of the bulk silicon via lapping – a coarse grinding process. Lapping is a machining process used across many industries to smooth a sample surface or remove a small amount of material using an abrasive. The abrasive can be fixed in place on one material *e.g.* sandpaper or emery cloth, or it can be free-flowing, *e.g.* an abrasive powder or slurry between two surfaces. Larger particle sizes result in a greater rate of material removal, thus care must be taken to choose an appropriate grit size to avoid lapping the die too quickly. Care must be taken to thoroughly wash the sample prior to lapping with successively finer

Table 4.4: FEPA sandpaper P-grades used in this investigation for lapping and their corresponding average particle sizes [20].

FEPA P-grade	Particle size range / μm
P600	25.8±1.0
P800	21.8±1.0
P1200	15.3±1.0
P2500	8.4±0.5

abrasive papers to avoid carrying over larger abrasive particles which will both ruin the surface finish and contaminate papers.

The lapping carried out in this investigation used a fixed-abrasive: a variety of wet-dry aluminium oxide (Al_2O_3) and silicon carbide (SiC) papers ranging from P600 to P2500 grit size. The Federation of European Producers of Abrasives (FEPA) specify the average particle size for each grade of sandpaper in different P-grades, as shown in Table 4.4.

When nearing the required thickness of silicon to be removed, it is vital to polish the surface using successively smaller grit sizes. These polishing steps are essential for data integrity, as the wet chemical etchant will selectively etch any scratches or pits down to the circuit layer far sooner than surrounding areas. The underlying circuit layer will be exposed by etching a scratch far earlier than surrounding areas; this over-etching would allow the chemical etchant an increased time to attack and dissolve the thin protective tunnel oxide layer while the rest of the silicon is still being etched, leading to the destruction of stored data.

The polishing conducted in this study involves free-flowing abrasive particles of diamond applied to a soft cloth polishing wheel. This is far less destructive than fixed-particle lapping due to smaller particle size and free-movement of abrasive particles. This step provides the final smoothing of the surface ready for chemical etching. The diamond pastes used were from the Buehler MetaDi range and available in sizes shown in Table 4.5.

As before, care must be taken to avoid cross contamination of various grit

Table 4.5: Buehler MetaDi diamond polishing paste particle sizes [21].

Paste Colour	Diamond particle size / μm
Deep Red	9
Yellow	6
Blue	1
Grey	0.25

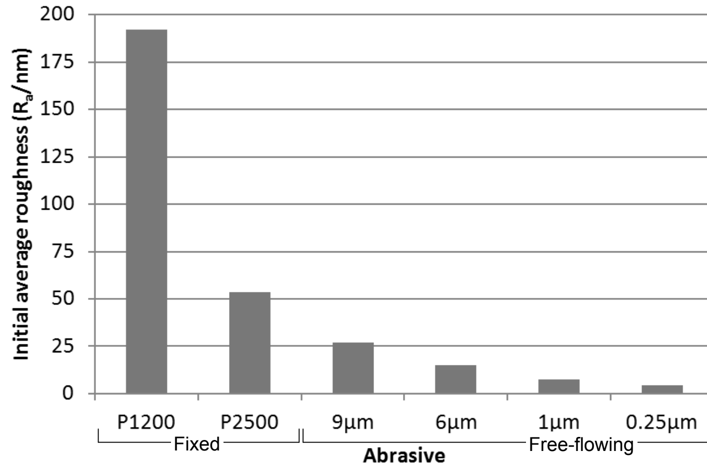


Figure 4.14: Surface finishes attained by lapping and polishing to varying stages with different abrasives.

sizes onto the sample or between polishing wheels. As the pastes are oil-based, mildly soapy water is recommended for cleaning the sample surface between successive grit sizes.

An investigation was carried out with the aim of assessing the effect of the lapped/polished silicon surface finish on the etch uniformity of TMAH. A silicon wafer was cut into square pieces of 5mm side length, similar in size to some of the larger SIM card microprocessors encountered. These pieces were set into resin using the method outlined above. Instead of directly lapping/polishing each sample to different surface finishes, samples were lapped (and polished where appropriate) to successively finer grades up to their designated final finish – replicating the mechanical bulk silicon removal process as accurately as possible. At this stage the 'initial' average roughness measurements were taken using a Veeco Dektak 6M profilometer, the results are shown in Fig. 4.14.

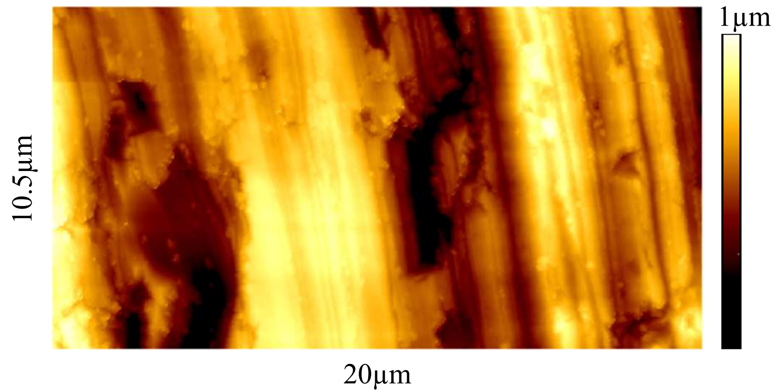


Figure 4.15: AFM topography scan of surface finishes after lapping with P1200 fixed abrasive.

The samples were given a brief dip in a buffered oxide etch solution to remove any natively grown SiO_2 layer, and then etched for 5 minutes in a solution of 25wt.% Aq. TMAH held at 70°C in a water bath. These chemical etchants form the core of the wet chemical etching processes used further along in the sample preparation method – more information on these are given later in this chapter. Various surface topographies resulting from lapping and polishing steps are shown in Figs. 4.15 to 4.17. After etching the roughnesses were measured once again, and the results are shown in Appendix A. Figure 4.18 shows a plot of the change in average roughness after etching against the initial average roughness. The results are clear, as smooth a finish as possible should be sought after. Obtaining a sufficiently low surface roughness will ensure a uniform etch profile during wet chemical etching. The final polishing step should be with a grit size no larger than 1μm.

To illustrate the effect that large scratches and pits in the surface topography have during wet chemical etching, a mounted microprocessor was etched with TMAH at a low temperature of 60°C and examined under an optical microscope every 20 minutes from 0–80 mins, Fig. 4.19. A similar effect can be see when etching with choline hydroxide solutions, shown in Fig. 4.20. Fig. 4.21 shows four optical micrographs of a partially etched processed sample, with scratches remaining from incomplete lapping/polishing being selectively etched

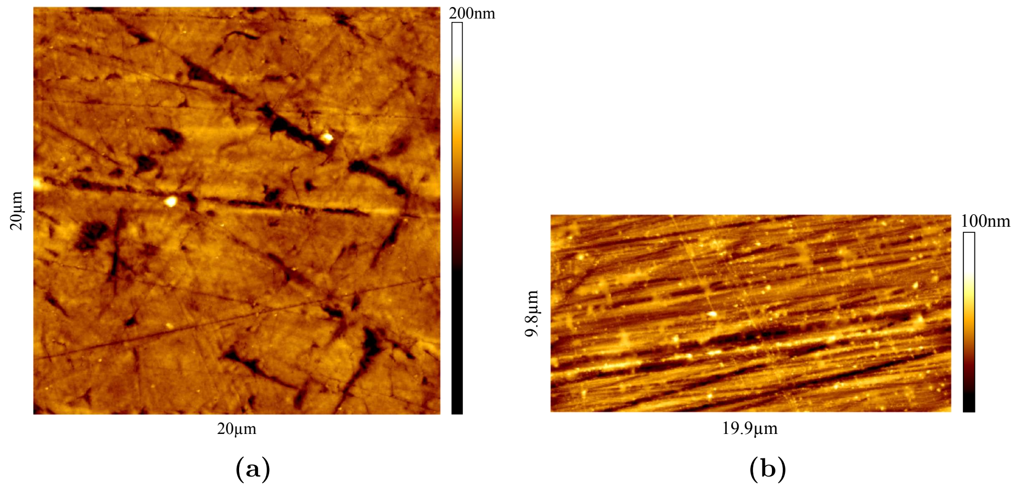
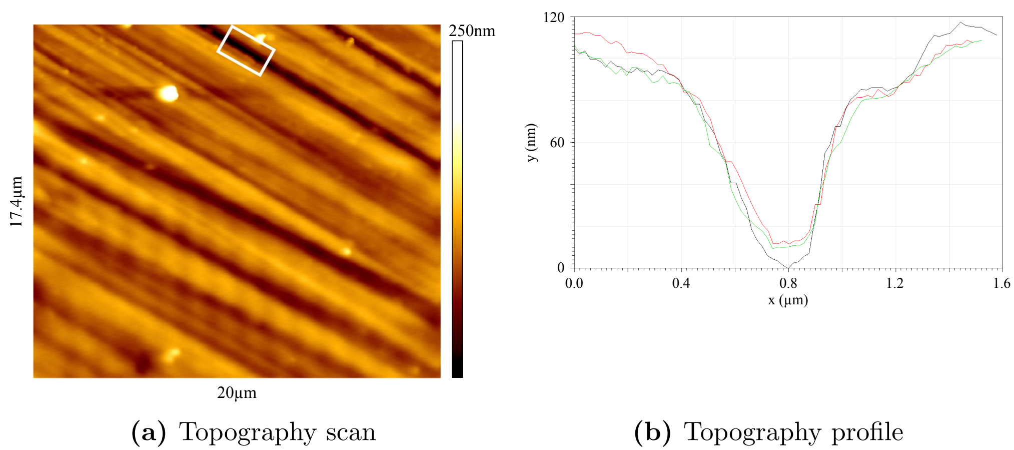


Figure 4.16: Two AFM topography scans of surface finishes after polishing down to 1µm diamond abrasive paste.



(a) Topography scan

(b) Topography profile

Figure 4.17: AFM topography scan **(a)** showing deep, near-parallel striations in the surface topography caused by lapping in one direction with fixed-particle abrasives taken after polishing down to 6µm diamond abrasive particles. Corresponding cross-sectional profiles at different points through the marked 100nm deep striation region **(b)**.

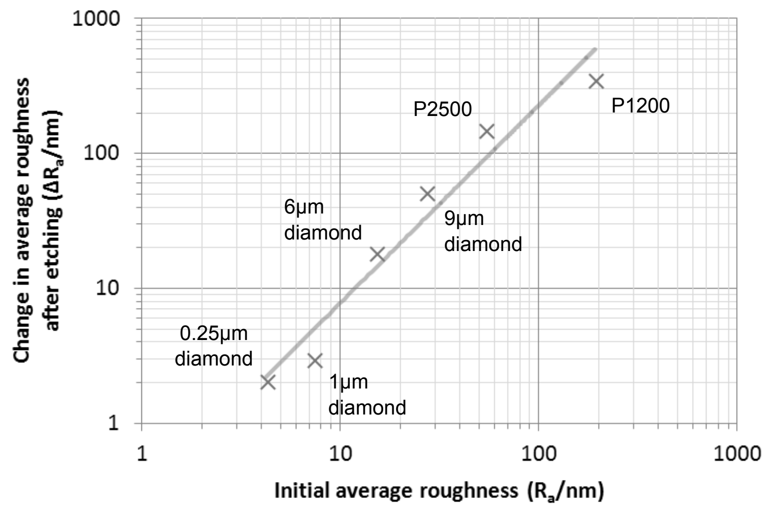


Figure 4.18: Plot showing the change in surface roughness from etching various initial surface finishes with TMAH 25wt% Aq. solution for 5 minutes at 70°C.

more rapidly than surrounding silicon – this prematurely exposes the circuitry beneath to hydroxide etchants, leading to destructive over-etching.

The final process determined to mechanically remove the majority of the bulk silicon begins by levelling off the topside of the resin 'puck' to remove any meniscus formed in the mould. Once this side is flat, the height of the puck can be ascertained using a drop-point micrometer. This allows the amount of material remaining at each stage to be calculated from the initial thickness of each die. The aim is to retain just under 40μm of silicon, enough for the selective chemical etchant to remove without endangering the thin tunnel oxides.

Most of the bulk silicon to be removed is lapped away carefully on a rotary water-lubricated grinding wheel using P600 or P800 wet-dry paper. If the sample is set at an angle, this is the stage to carefully correct this issue. After the required bulk material has been removed, progressively finer grade papers are used to clean up the surface, smoothing it more and more while removing silicon at a progressively lower rate. After the finest grade paper available, P2500 in this study, has been used and optical inspection under a microscope reveals no

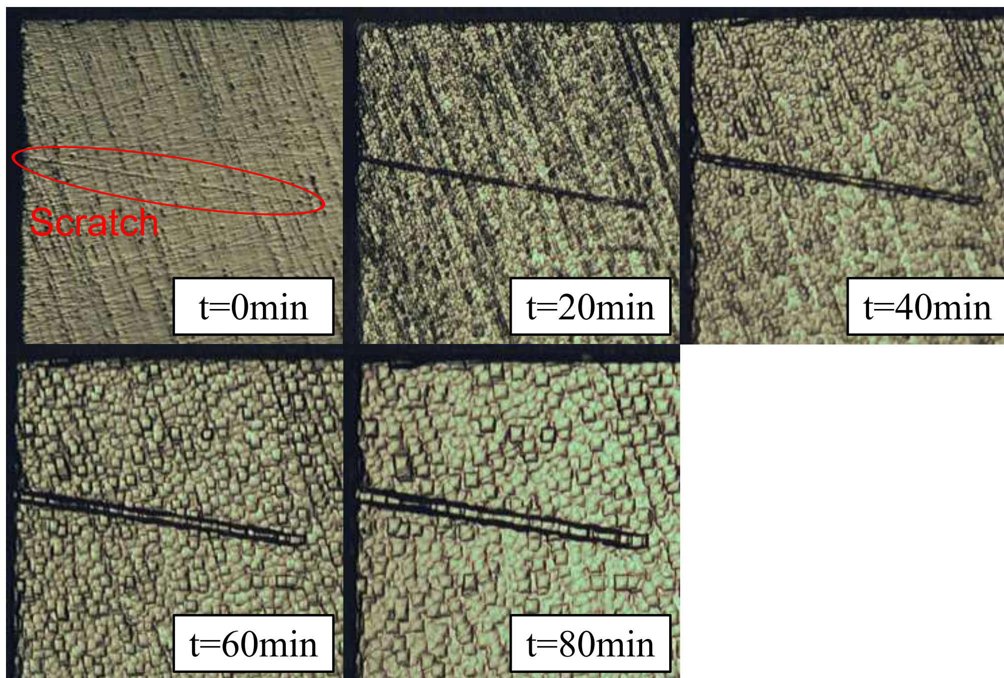


Figure 4.19: The effect of surface scratches on the wet chemical etching process. Etching was conducted at 60°C with TMAH 25wt% Aq. solution. The initial scratch has been highlighted at t=0 mins, and remains a persistent feature throughout the 80 minute etching process.

Figure 4.20: The effect of surface scratches on the choline hydroxide etching. 2.5× optical zoom. From left-right: 0 mins, 40 mins, 65 mins. From [8].

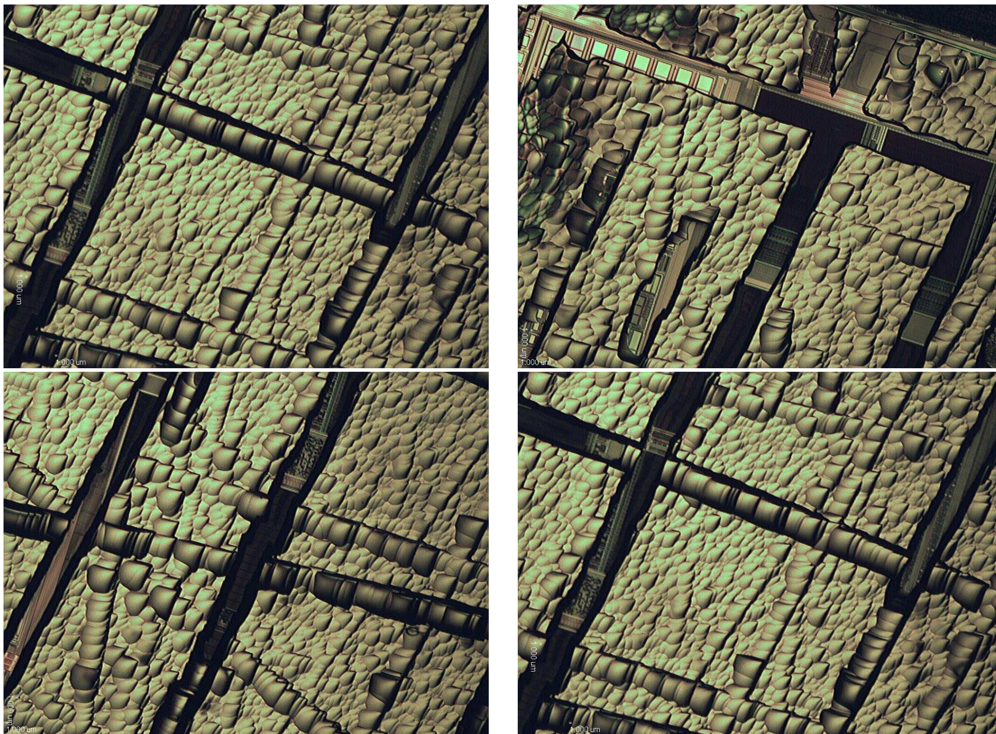


Figure 4.21: Four optical micrographs of a partially etched processed sample illustrating how scratches left from lapping are selectively etched down to the circuit layer far sooner than surrounding silicon, images captured at t≈55 mins. Etching conducted with TMAH 25wt% Aq. at 90°C.

outstanding scratches or pits on the surface, polishing can begin.

As with abrasive papers, progressively finer particle-size diamond pastes are used on their designated felt polishing wheels. These wheels must be kept separate to avoid cross-contamination, and the sample should be gently washed between each polishing step with mildly soapy water. The results shown above indicate that it is worthwhile polishing to a minimum of 1μm diamond grit size, but below this it seems to make little difference to etching uniformity (however, if available, they should be used).

4.5 Wet chemical etching

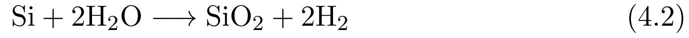
With the sample's bulk silicon now thinned to below 40μm, the next step is to complete the removal using chemical etching to safely expose the underside of the circuitry ready for SPM examination. For this etch to work, a sufficiently selective etchant must be chosen. Use of plasma etching is inadvisable given the charged ions and the disruptive effect they would have on the stored data. Before a selective chemical etchant can be used on silicon, the native oxide layer must first be removed with a buffered solution of hydrofluoric acid.

4.5.1 Silicon dioxide removal

Silicon dioxide will grow on a bare silicon surface when exposed to oxygen or water in so-called 'dry' and 'wet' oxidation respectively; see Eqs. (4.1) and (4.2). This native oxide growth is initially a surface reaction, but as the SiO_2 builds up the incoming oxygen diffuses through the oxide layer to reach the silicon interface. The Gibbs free energy[2], ΔG, at 25°C for these reactions is −857kJ/mol for dry oxidation and −382kJ/mol for wet oxidation [22].

[2] $\Delta G < 0$ denotes a favoured reaction, $\Delta G = 0$ denotes equilibrium, $\Delta G > 0$ denotes an disfavoured reaction

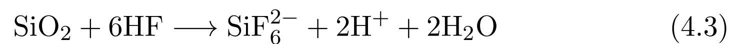
$$\mathrm{Si} + \mathrm{O_2} \longrightarrow \mathrm{SiO_2} \tag{4.1}$$



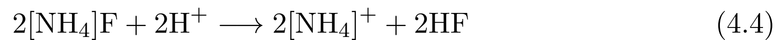
$$\mathrm{Si} + 2\mathrm{H_2O} \longrightarrow \mathrm{SiO_2} + 2\mathrm{H_2} \tag{4.2}$$

Oxygen will eventually have insufficient energy to penetrate the thickness of the oxide layer to continue oxide growth – reactions between silicon and oxygen cease at a thickness around 25Å, though this range can vary between 12–30Å [23–25]. Although a very thin layer, 25Åis still sufficient to inhibit silicon etching by up to 15 minutes, and if uneven, could result in a non-uniform etch profile.

Prior to any silicon etching with hydroxide solutions, any native SiO_2 must be removed. The buffered oxide etch (BOE, sometimes called BHF) used was a pre-mixed solution of 40% aqueous ammonium fluoride ($[NH_4]F$) solution and 49% hydrogen fluoride (HF) aqueous solution mixed to a 6:1 ratio by volume. The resulting mixture etches SiO_2 at a rate of approximately 2nm/s at 25°C. It is commonly used in the semiconductor industry as a means of removing any silicon dioxide that has formed prior to deposition of other materials.



$$\mathrm{SiO_2} + 6\mathrm{HF} \longrightarrow \mathrm{SiF_6^{2-}} + 2\mathrm{H^+} + 2\mathrm{H_2O} \tag{4.3}$$



$$2[\mathrm{NH_4}]\mathrm{F} + 2\mathrm{H^+} \longrightarrow 2[\mathrm{NH_4}]^+ + 2\mathrm{HF} \tag{4.4}$$

$$2[\mathrm{NH_4}]^+ + \mathrm{SiF_6^{2-}} \longrightarrow [\mathrm{NH_4}]_2\mathrm{SiF_6} \tag{4.5}$$

HF reacts with silicon dioxide to form hexafluorosilicate ions (SiF_6^{2-}), hydrogen ions and water, as shown in Eq. (4.3). The ammonium fluoride in solution acts as both a buffer to maintain HF concentration, see Eq. (4.4), and a source of ammonium ions to form ammonium hexafluorosilicate ($[NH_4]_2SiF_6$), see Eq. (4.5), which readily dissolves into solution. BOE is an essential tool for removing any native oxide ready for silicon etching.

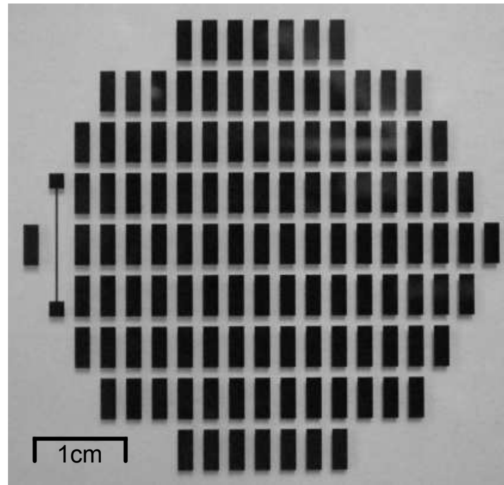


Figure 4.22: Photograph of the 2 inch photolithographic mask used to produce samples for chemical etch rate tests.

Characterising the BOE etch process

The typical method to tell if the BOE is finished is to observe the hydrophilic/ hydrophobic properties of the sample surface; silicon is hydrophobic, whereas silicon dioxide is hydrophilic. This technique, while crude, is extremely effective because even a thin oxide layer present on the surface will exhibit hydrophilia – staying wet when removed from BOE; but once etched down to silicon the solution will bead and run off the surface, exhibiting hydrophobia.

To ascertain the etch rate of BOE, samples were fabricated from a silicon wafer with a 3µm thermally-grown oxide layer. The wafer had a film of positive photoresist spun onto the surface, and was patterned with the photolithography mask shown in Fig. 4.22. The clear (developed) areas between masked rectangles were etched with HF to create oxide mesas. The mesas varied between 2.98–3.04µm, measured with an Alphastep 200 step profiler. For safety reasons, it should be noted that the HF dip setup was fixed inside a fume cupboard with no means of agitation or heating, the solution remained at approximately 20°C throughout.

The wafer was cut into pieces and etched in the pre-prepared BOE etching

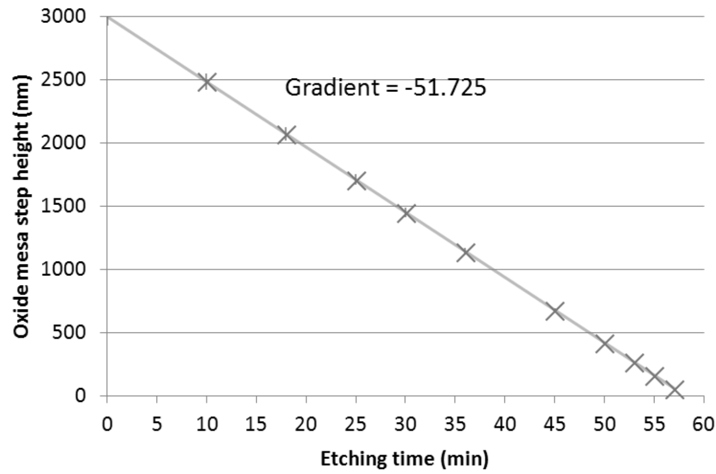


Figure 4.23: Oxide mesa step heights throughout prolonged buffered hydrofluoric acid (BOE/BHF) etch. The calculated gradient of the fitted line gives an etch rate of 51.7nm/min.

solution in the cleanroom, the pieces were removed, washed, and the step height was checked on the step profiler every few minutes, the data are shown in Appendix B. By 58 minutes two of the areas had been fully etched and exhibited hydrophobic properties, but one area of oxide remained – etched away within the next minute. Once all the oxide mesas had been fully etched away, the remaining step produced from BOE etching silicon for the duration of the test was negligible; estimated to be <2nm this gives a silicon etch rate of 0.03nm/min. The results are shown graphically in Fig. 4.23 – the gradient of the curve shows the etch rate under these conditions to be 51.7nm/min, with an etch selectivity of approximately 1500:1 (SiO_2:Si).

4.5.2 Silicon removal

Various etchants exist for etching silicon, each of which has advantages and disadvantages which must be considered carefully according to the application. The main requirement for silicon etching in this study is a selectivity of Si:SiO_2 large enough to etch away the silicon underside of the die without harming the thin tunnel oxide layer. The higher the selectivity, the more freedom afforded in

the mechanical silicon removal steps, and thus a greater degree of safety afforded the samples.

Etchants such as ethylenediamine pyrocatechol (EDP) were avoided due to its extreme corrosive and carcinogenic nature and the more complex etching setup required. EDP must be used in a highly controlled environment, with a reflux condenser to keep the volatile components in solution. It will readily oxidise any nearby metals and is noted for its ability to etch aluminium. By far the most popular method for selectively etching silicon is to use potassium hydroxide (KOH).

Silicon removal: potassium hydroxide

With an etch selectivity for Si:SiO_2 of 500:1, KOH offers a fast and reliable method of etching silicon while leaving silica relatively untouched. Silicon etches consisting of a hydroxide solution work by introducing excess hydroxyl anions necessary for the formation of a water-soluble silicate complex. This etching process has been widely studied, with researchers such as Zubel and Tanaka leading the field.

Zubel has published numerous papers covering KOH etching. Her paper on the development of etch hillocks [26] covers many performance parameters of wet chemical etching with hydroxide solutions and presents a growth model of mesa structures and hillocks. Incomplete dissolution of reaction products, foreign impurities such as metals, and hydrogen bubbles residing on the surface can all lead to the formation of etch hillocks through micro-masking. They make various recommendations: agitation with a stirrer or, if possible, an ultrasonic bath to remove hydrogen bubbles and reaction by-products; addition of IPA [27–30] to reduce surface tension, facilitating the removal of hydrogen bubbles reaction products; using a higher concentration of hydroxide solution; and the addition of an oxidising agent, for example oxygen saturation to help remove hydrogen.

The addition of IPA seems to be a trend throughout hydroxide etching pro-

cesses. Models explaining the effects of IPA on the etching process in various planes are given in [31, 32] and in far greater detail in [33]. The models explain hillock formation and how IPA helps to even out surface morphology. However, in Si(100) orientation the IPA competes with the hydroxide ions for surface binding sites, but despite a dramatic improvement in morphology, it is still unknown how it interacts. In the case of etching the silicon backside of SIM card microprocessors, a more uniform etch rate is achieved by including IPA in the silicon etch solution. This helps preserve data integrity by reducing selective erosion where scratches remain after polishing.

Tanaka *et al.* published two studies [34, 35] into the effects of impurities at a level of parts per billion (ppb). It appears that even a low level of lead or copper contamination will lower the etch rate by altering the etching chemistry. Copper impurity above 100ppb in KOH solution will dramatically affect the roughness of the sample through copper metal deposition onto the surface forming micromasks. Tanaka *et al.* also published a paper detailing their investigation into fast etching with KOH. They managed to achieve smooth surface etching at up to 20μm/min for Si(100) and 10μm/min for Si(110) near the boiling point of 50wt.% of KOH (around 150°C).

Despite being a popular silicon etchant, an insufficient etching selectivity of 500:1 will not ensure that the tunnel oxide remains intact through etching, and the inevitable introduction of alkali potassium ions makes it incompatible with CMOS processing as required in the removal of bulk silicon to gain access to the EEPROM array. These ions could also interfere with SKPM measurements performed in electrical examination of the memory array.

Silicon removal: ammonium hydroxide

Ammonium hydroxide is an effective silicon etchant and does not introduce alkali metal ions. The problem, however, is that at the higher temperatures required to attain a sufficient etch rate, the ammonia would evaporate rapidly from the

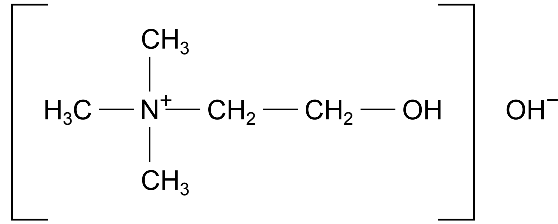
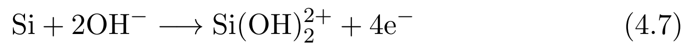


Figure 4.24: Chemical structure of choline hydroxide.

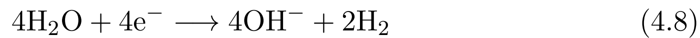
solution. Thus it is common practice in semiconductor processing to use a variant whereby the ammonium hydroxide is ballasted with a less volatile organic – tetramethylammonium hydroxide (TMAH) or choline hydroxide are two of the more commonly used silicon etchants.

The chemistry behind hydroxide etching of silicon is as follows: the hydroxyl anions released by ammonium hydroxide (Eq. (4.6)) oxidise the silicon to form a silicate (Eq. (4.7)), which further reacts with hydroxyls from the reduction of water (Eq. (4.8)) to form a water-soluble silicate complex (Eq. (4.9)).

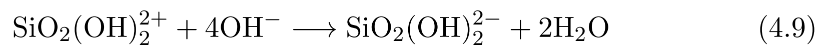
$$2(CH_3)_4NOH \longrightarrow 2(CH_3)_4N^+ + 2OH^- \tag{4.6}$$



$$Si + 2OH^- \longrightarrow Si(OH)_2^{2+} + 4e^- \tag{4.7}$$



$$4H_2O + 4e^- \longrightarrow 4OH^- + 2H_2 \tag{4.8}$$



$$SiO_2(OH)_2^{2+} + 4OH^- \longrightarrow SiO_2(OH)_2^{2-} + 2H_2O \tag{4.9}$$

Silicon removal: choline hydroxide

One possible ballasted hydroxide etchant is choline hydroxide, see Fig. 4.24. Choline hydroxide's advantage over the more widely-used and less toxic TMAH is a slightly higher silicon to silicon dioxide etch selectivity of around 5000:1.

Choline hydroxide's use as an alternative to TMAH in failure analysis applications is discussed in greater depth by Korchnoy [36]. This paper discusses the requirements for CMOS failure analysis, such as an etchant's selectivity to keep oxide layers intact, and similar etch rate dependence on doping type and level to

keep etching constant across various structures. A solution of choline hydroxide 50wt% Aq. was used throughout her work, with various studies conducted.

For the required application in this study, the major disadvantage of etching with choline hydroxide is the cleaning required afterwards. TMAH can be cleaned from a surface by immersion in DI water and drying with a nitrogen gun; however, choline hydroxide requires a far more extensive cleaning process due in part to its viscosity. Korchnoy outlines an extensive cleaning procedure for samples after etching with choline hydroxide: an initial quench and soak in DI water for several minutes, an 8 minute soak in boiling acetone, followed by a 2 minutes of ultrasonic cleaning in acetone, 6 minutes cleaning in boiling methyl alcohol, followed by a 2 minute ultrasonic clean also in methyl alcohol, and finally concluding with a third ultrasonic clean in isopropyl alcohol for 2 minutes at room temperature.

After the first few batches of SIM cards had been etched with TMAH, it became evident that etching with choline hydroxide was out of the question. The extreme delicacy of the samples post-etching would definitely not withstand the ultra-rigorous cleaning process required for choline hydroxide, TMAH continued to be used as the silicon etchant.

4.5.3 Silicon removal: tetramethylammonium hydroxide

Tetramethylammonium hydroxide (TMAH), see Fig. 4.25, is now commonplace in the realm of silicon etching as an alternative to KOH. Advantages over KOH include a much greater selectivity, lower toxicity, and no metal ions to contaminate the sample surface during etching – making it ideal for industrial CMOS processes. TMAH is commonly available in solution with water, and less commonly methanol.

Various solution temperatures have been used among TMAH-related publications; Liebert's investigations [1, 19] into *Failure analysis from the backside of a die* and *Encapsulation of naked dies for bulk silicon etching with TMAH*

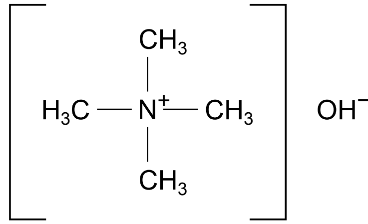


Figure 4.25: Chemical structure of tetramethylammonium hydroxide (TMAH).

both use TMAH 25% Aq. solution heated to 105°C for etching, close to the boiling point of 109°C, but it appears far more common to use solutions in the range 80±10°C. Silicon etch rates are well documented and generally agreed upon [27, 37–40], when taking into account differences in experimental methods *e.g.* agitation, concentration changes over time and with re-use, temperature stability.

With an etch selectivity for Si:SiO_2 of greater than 4000:1 [1], TMAH is a more desirable anisotropic silicon etchant than KOH (500:1 selectivity). It also has the advantage of flat etch profiles, where KOH produces a V-shaped etch profile [41]. Tabata *et al.* [39] showed that the Si(100) etch rate of TMAH solutions increases as the concentration decreases, much like KOH, and further that smooth silicon surfaces can only be obtained for solutions above 22%. Chen *et al.* [42] also reported similar findings at 90°C, that the roughness of TMAH solutions increased dramatically below 10%, with maximum reported R_a >800nm. The increase in roughness appears to follow a short term exponential trend, but more detailed examination of the data is required to confirm this. Chen explicitly states that a concentration greater than 15% is required for a surface roughness below 20nm.

The addition of IPA into TMAH solutions is discussed by Zubel [27], beginning with a simple outline of the silicon etching process. It is suggested that the tetramethylammonium (TMA^+) ions adsorb onto the silicon surface, blocking the hydroxide (OH^-) ions from etching – thus higher concentrations of TMAH

actually decreases the etch rate. An accepted model of the silicon etching process is given in 5 steps:

1. Diffusion of reagent particles towards silicon surface
2. Adsorption of reactive and non-reactive particles (ions) on silicon surface
3. Surficial reactions (oxidation of silicon)
4. Desorption of reaction products
5. Diffusion of reaction products from silicon surface to the bulk of solution

When added to TMAH solutions, IPA plays a similar role to TMA^{+} ions, acting to decrease the etch rate. In planes other than (100), the TMA^{+} ions do not adsorb onto the surface, thus with increased concentration, an increase in etch rate is observed. This has a smoothing effect as any errant morphological features that form on the surface during etching, which inherently contain facets other than (100), are more rapidly etched away, bringing the surface back towards a (100) planar morphology.

To further explain the smoothing effect that IPA has on hydroxide etching, the equilibrium conditions between etching and desorption rates must be considered. An etch rate higher than the desorption rate leaves an excess of reaction by-products on the surface. These act to hinder further reactions, locally masking that area from oxidation until the reaction by-products can be removed – this leads to etch roughened surfaces. If, however, the two rates are in equilibrium, then the surface is cleared of by-products at the same pace as etching occurs. This results in a uniform etch profile and produces smoothly etched surfaces. The effect of the addition of IPA is thus twofold: it reduces the surface tension, facilitating the removal of etch by-products and bubbles; and it increases the number of competing species trying to adsorb onto the silicon surface, thus lowering the rate of silicon removal.

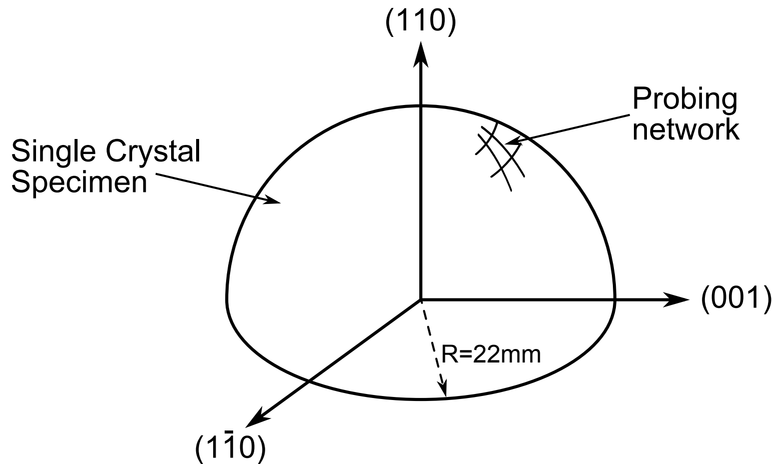


Figure 4.26: Silicon hemisphere used in surface etching studies by Sato [41], Shikida [43, 44] & van Veenendaal [45, 46].

Chen *et al.* [42] also offers a similar model with three factors affecting etch rate: reactant transport to the surface, surface reaction, and the transport of products away from surface. If the first or last are the limiting factors, then the etch rate is diffusion-limited and can be increased by agitating the etchant solution. Diffusion-limited processes are indifferent to temperature variations as they have lower activation energies than temperature-dependent reaction rate processes.

The various possible morphological surface features generated by wet chemical etching different silicon planes at different TMAH concentrations are shown excellently by Kramkowska and Zubel [47], of particular note the smoothing effect shown in planes other than (100) when comparing TMAH 25% against TMAH 20% + IPA 10%. While the smoothest surfaces obtained were from TMAH 10% + IPA 20%, this changed the etching characteristics of the solution from anisotropic to "almost isotropic".

Also of note are papers by van Veenendaal [45, 46], Sato [41] and Shikida [43, 44] where a silicon hemisphere, see Fig. 4.26, of radius 22mm was used to assess the effects of etching different planes of silicon. Features from triangular and hexagonal etch pits on (111), spherical depressions on (100) and zigzag

Table 4.6: Etch rates measured from hemispherical sample. After [43].

Crystallographic Orientation Si(*hkl*)	TMAH Etching Rate (nm/min) (20.0wt.%, 79.8°C)
100	603
110	1114
210	1154
211	1132
221	1142
310	1184
311	1223
320	1211
331	1099
530	1097
540	1135
111	17

patterns on (110) were all documented with optical microscopy. The orientation-dependent etching rates measured in their studies are shown in Table 4.6.

To avoid transient temperature change with a sample of such a large mass, it was heated in a dry container inside the water bath prior to immersion into TMAH. When the specimen reached the etching temperature, it was immersed into 1 litre of TMAH solution and held at least 5mm from the sides of a Teflon basket. Fresh etchant was used in every experiment, and a magnetic stirrer was used in the water bath to help equalise any temperature gradients. No means of circulation were used in the etchant solutions. By using this method, a temperature stability of ±0.9°C was achieved with uniform distribution.

The average surface roughness of the (100) orientation was around 0.9µm at 10wt.%, but this stabilised to around 0.4µm above 20wt.%. Of particular note in their work is the statement that unlike etching with KOH, etchant circulation when using TMAH must not be ignored – this relates to the more difficult to attain equilibrium condition and hints at etchant circulation as a means of facilitating etch by-product removal.

Developing the TMAH silicon etch process

A series of experiments were conducted to characterise the TMAH silicon etch process. The aim is to use TMAH to safely remove the remaining bulk silicon, etching down to the tunnel oxide layer of the deposited microprocessor circuitry. The primary concern at this point is achieving a uniform etch profile as TMAH is known to selectively etch scratches and pits faster than surrounding areas, potentially putting the tunnel oxide integrity at risk.

Samples similar to those fabricated for the BOE etch rate tests, see Section 4.5.1, were used to also gauge the etch rate for TMAH. These samples consisted of 3μm rectangular oxide mesas surrounded by silicon trenches. The silicon trench initial arithmetic average surface roughness (R_a) was measured to be consistently below 1nm (ranging from 0.42 to 0.86).

Solutions of TMAH 25wt% Aq. and TMAH 25wt% Aq. with IPA 10vol% were assessed. The two methods of heating the etching solutions were a hotplate and a water bath. The hotplate had the option of a magnetic stirrer to help reduce the temperature gradient throughout solution, while the water bath offered no means of agitation. Solutions heated in the water bath had a greater temperature stability, ±0.5°C, while those heated with a hotplate fluctuated ±3°C. A schematic of the hotplate setup is shown in Fig. 4.27.

The setup using a water bath to heat solutions was very similar, albeit without a magnetic stirring bar providing agitation – the only agitation was thus due to the natural circulation provided by the rising hydrogen bubbles from etching. In either case, the sample must be kept away from the bottom and sides of the etching container. A PTFE basket with large holes was used to this effect, allowing the samples to be held near the centre of the etching solution at all times.

The samples were placed into etching solutions held at specific temperatures for 5 minutes each, with temperatures monitored using a K-type thermocouple.

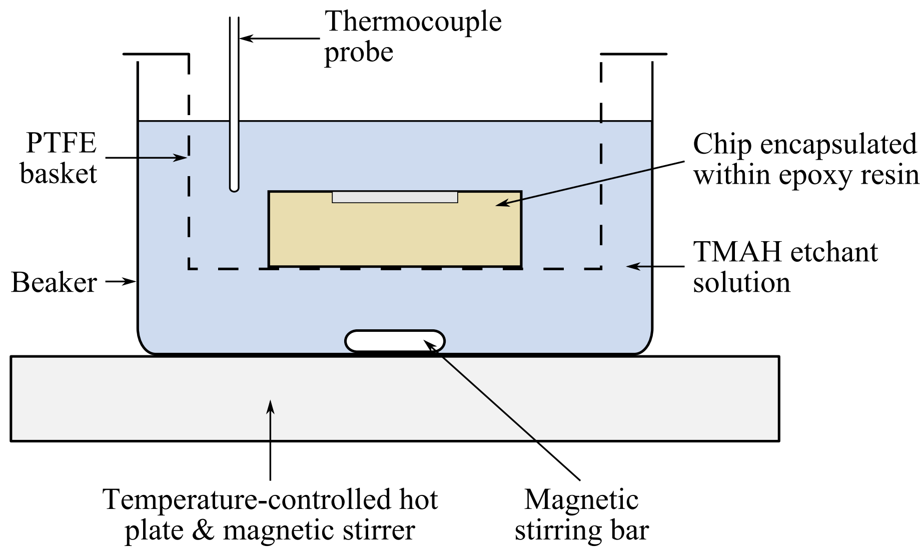


Figure 4.27: Schematic showing setup used for etching on a hotplate.

Recalling the 4000:1 Si:SiO_2 selectivity, TMAH will barely affect the oxide mesa, but will actively etch the silicon trenches deeper. Upon removal, the samples were quenched in DI water, rinsed in acetone, and blown dry with a nitrogen gas gun.

Once cleaned, the trench's etched surface roughness was once again measured. This step was carried out prior to BOE etching to avoid altering the surface morphology during the prolonged oxide etch. After measuring the surface roughness, a long oxide etch in BOE is performed to remove the oxide layer and reveal the silicon step created by TMAH etching the trench regions. The data gathered from these experiments are shown in Appendix C.

The Si(100) etching rates over a range of temperatures for solutions of TMAH 25wt% Aq. with and without agitation is shown in Fig. 4.28. Agitating the etching solution results in a linear relationship between temperature and etch rate. At lower temperatures tested, the non-agitated solution followed this trend, exhibiting similar etching rates, however, the curve begins to level out at temperatures above 80°C. This flattening signifies a reduction in the rate of change of material dissolution with increasing temperature, and could be explained by

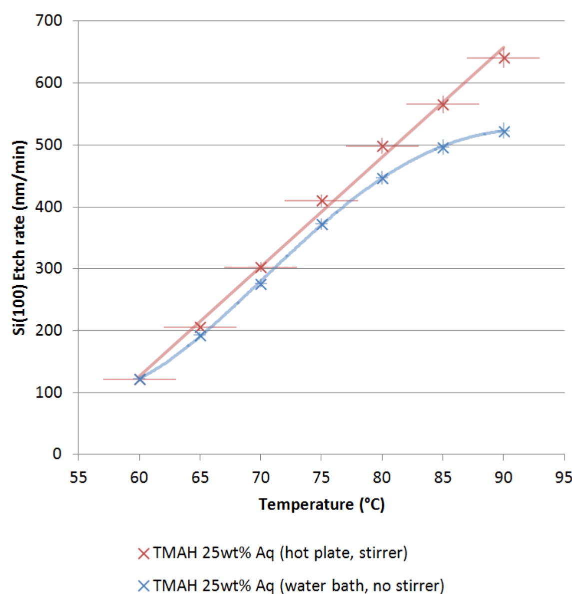
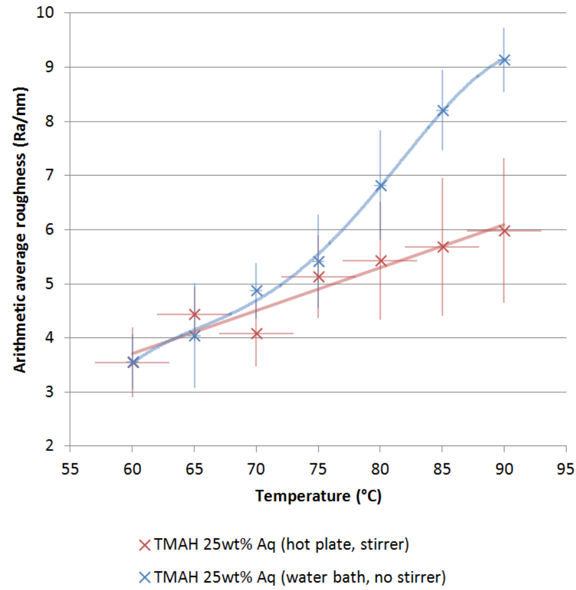


Figure 4.28: Graph showing Si(100) etch rates for TMAH 25wt% Aq. solutions with and without agitation over a range of temperatures.

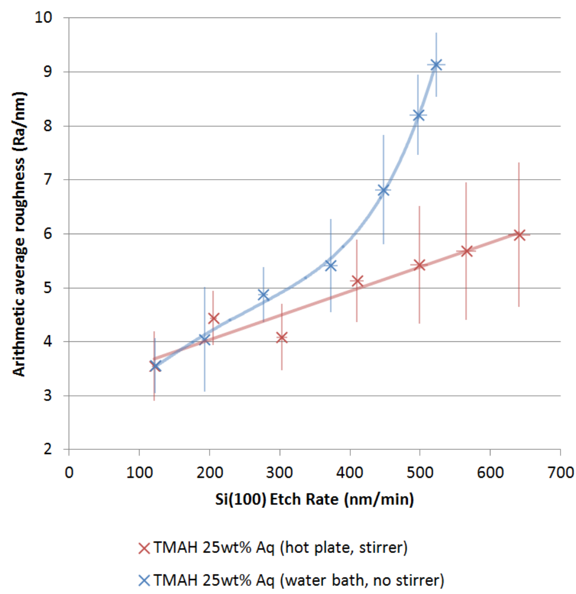
shifts in the previously discussed equilibrium conditions between etching and desorption rates. If the solution is not agitated sufficiently, etch by-products and bubbles cannot desorb from the silicon surface at the same rate at which etching occurs, slowing the etch rate significantly, and micromasking areas of the surface. The fact that introducing additional agitation removes this effect indicates that without agitation, etching with this solution is a diffusion-limited process at higher temperatures.

Etching with TMAH solutions without agitation shows a notably higher roughness than etching with agitation at temperatures above 75°C, see Fig. 4.29a. A plot of the surface roughness against the calculated Si(100) etch rates is shown in Fig. 4.29b. This rise coincides with the levelling off of the etch rate seen in graph Fig. 4.28, backing the concept of increased micromask production outside of equilibrium conditions.

A comparison of etch rates with varying temperature for agitated solutions of TMAH 25wt% Aq. and TMAH 25wt% Aq. with IPA 10vol% is shown in Fig. 4.30. The addition of isopropanol into the etching solution clearly has the



(a) Average roughness vs. etchant temperature



(b) Average roughness vs. Si(100) etch rate

Figure 4.29: Graphs showing average roughnesses for TMAH 25wt% Aq. solutions with and without agitation.

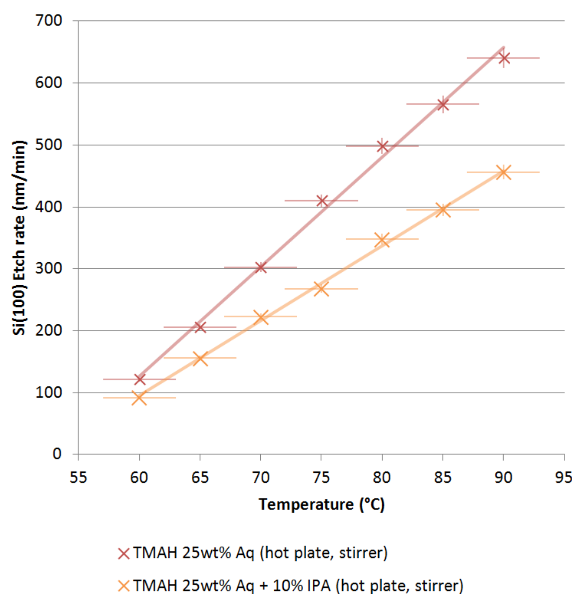
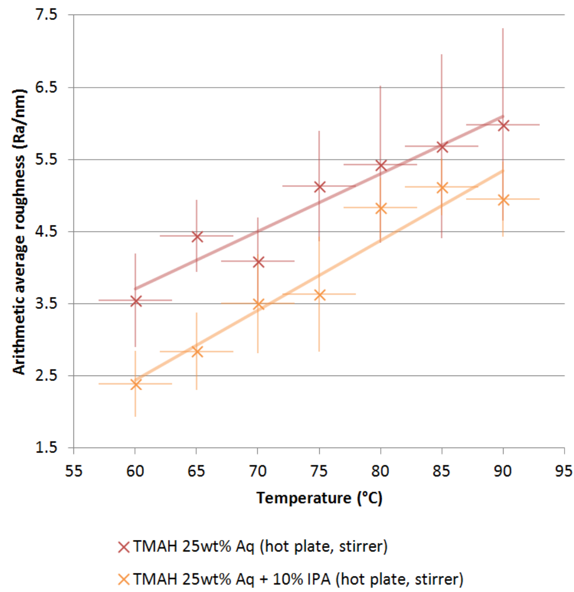


Figure 4.30: Graph showing Si(100) etch rates for TMAH 25wt% Aq. and TMAH 25wt% Aq. with IPA 10vol% solutions with agitation over a range of temperatures.

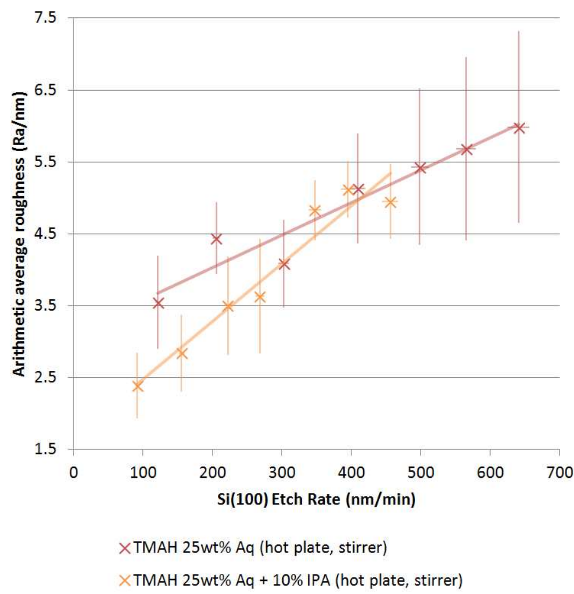
effect of lowering the Si(100) etch rate as well as the rate of change of etching with temperature. While the inclusion of IPA doesn't affect the rate of reaction for hydroxide anions, it does introduce other molecules which vie for surface adsorption against the reactant ions, effectively reducing the rate of reaction.

The addition of IPA into the etching solution produces a notably lower surface roughness across all temperatures, this can be seen in Fig. 4.31b. However, given the shift in etch rates upon inclusion of IPA, a comparison between the calculated etch rate and resulting surface roughness better demonstrates the surface smoothing than etch temperature, this is shown in Fig. 4.31a.

Unlike TMAH-only solutions, agitating etching solutions containing IPA 10vol% has no substantial effect on the rate of silicon removal. The similar etch rates for solutions of TMAH 25wt% Aq. with IPA 10vol% at varying temperatures with and without agitation can be seen in Fig. 4.32. The two diffusion-limiting factors are the rate of reactant surface adsorption, and the rate of reaction by-product desorption. IPA acts to reduce both of these, first by introducing



(a) Average roughness vs. etchant temperature



(b) Average roughness vs. Si(100) etch rate

Figure 4.31: Graphs showing average roughnesses for TMAH 25wt% Aq. and TMAH 25wt% Aq. with IPA 10vol% solutions with agitation.

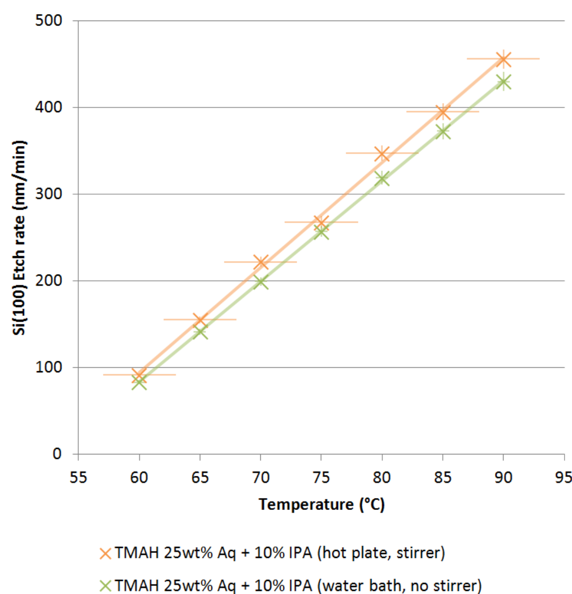


Figure 4.32: Graph showing Si(100) etch rates for TMAH 25wt% Aq. with IPA 10vol% solutions with and without agitation over a range of temperatures.

additional ions to vie for surface adsorption against hydroxide anions, and second by lowering the surface tension in solution, allowing more rapid desorption of reaction by-products from the surface, and easier release of hydrogen bubbles. From this it is possible to surmise that the equilibrium conditions for this solution have been altered sufficiently by the inclusion of IPA alone to produce a temperature-limited process, and that further diffusion-increasing factors, *i.e.* agitation, have no effect.

Comparing solutions containing IPA with varied presence of agitation, see Fig. 4.33, it is clearly visible that in both cases the surface roughness rises with increasing temperature/etch rate. Factoring in error boundaries for the data set without agitation, the increasing rate of change of surface roughness could be seen to be almost linear, especially when plotted next to similar data set produced with agitation. In either case, their close proximity to one another illustrates the effectiveness of the addition of IPA at reducing errant surface morphological features, *e.g.* hillocks, from forming, presenting a smoother etch

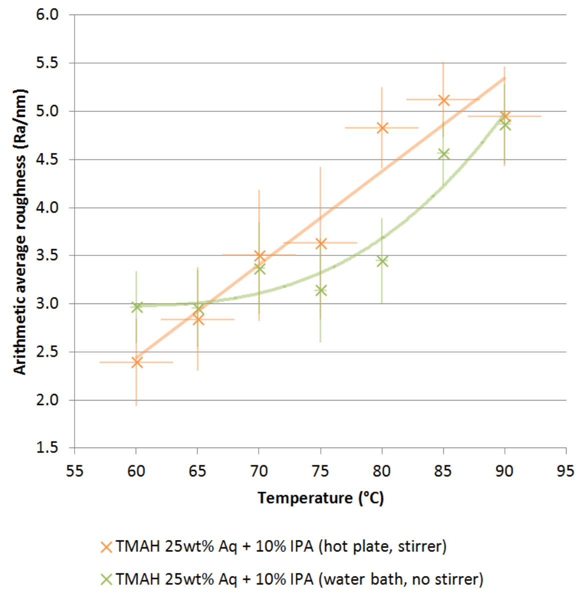
profile even without additional agitation.

From these results the etching process developed to remove the remaining backside silicon is to use a solution of TMAH 25wt% Aq. with IPA 10vol% heated in a water bath to a temperature not exceeding 80°C. This should provide a Si(100) etch rate of approximately 320nm/min (19.2µm/hr) at 80°C. At this rate, the remaining 40µm of bulk silicon left after lapping and polishing should be etched away in approximately 2 hours. To be safe, the sample should be checked for the first appearance of circuitry every 15 minutes from 75 mins, and once sighted every 5 minutes thereafter for completion.

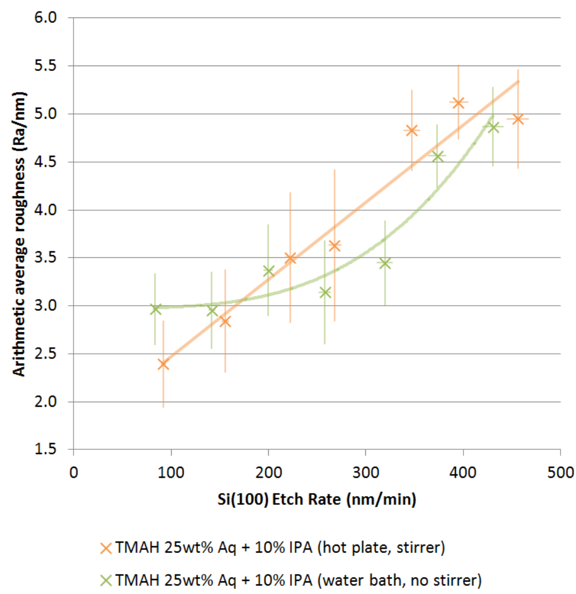
If a water bath is unavailable, a hotplate with magnetic stirrer will suffice, but the temperature should be set no higher than 75°C to avoid overshooting 80°C during heating bursts; this will result in an etch rate of approximately 270nm/min (16.2µm/hr). With the addition of IPA, solution circulation is less of a concern than without, but the temperature gradient from heating with a hotplate should be considered. With this in mind, if a magnetic or overhead mechanical stirrer is available to aid in solution circulation for a water bath setup then this is not to be overlooked, providing a secondary means of ensuring a temperature-limited rather than a diffusion-limited process.

Throughout etching it is imperative to keep the face of the resin mould containing the embedded die facing upwards and away from the sides of the etching container. Use of a Teflon basket proved excellent in this regard, able to keep the sample in the centre of the solution throughout. The rinse process should be conducted as follows: a gentle quench in warm (≈40°C) DI water for a few minutes, then the same again at room temperature, followed by a 2 minute dip in acetone, the same in IPA, and finally very gently blown dry with nitrogen.

To achieve a consistent etch rate it is important to use fresh etchant solution each time – re-use of etching solutions is a common practice in many fields, but should be avoided to ensure adherence to a forensically-sound process. It was found in earlier tests that the concentration and effectiveness of these solu-



(a) Average roughness vs. etchant temperature



(b) Average roughness vs. Si(100) etch rate

Figure 4.33: Graphs showing average roughnesses for TMAH 25wt% Aq. with IPA 10vol% solutions with and without agitation.

tions can vary substantially depending on their history. As the concentrations of hydroxide solutions decrease with use, fewer ions are present to compete with reactant hydroxide ions, thus resulting in a substantially increased etch rate. While early test samples with smooth surfaces were less affected, lapped samples exhibited a more dramatic effect – any scratches or pitted regions resulting from lapping experienced the full effect of being preferentially etched at a far higher rate, forming deep trenches and ultimately resulting in a dangerously non-uniform etch profile. The effects were similar to those shown earlier in Figs. 4.19 and 4.20.

4.6 Summary of preparation process

The final process developed to prepare SIM card samples for circuit underside SPM examination is shown in Fig. 4.34, and described as follows:

An initial visual inspection should be conducted on all samples to check for external damage. Particular attention should be paid to the contact pad area and the region behind the chip module. The main card body should be separated from the chip module by placing on a hotplate heated to between 120–150°C for 3–5 seconds. With the hot melt adhesive softened, the two parts can be carefully pulled apart using pairs of tweezers. At this point a second optical inspection of chip module should be conducted to assess any damage to the bond wires or the chip itself.

The chip module should be suspended in a small beaker of fuming nitric acid heated on a hotplate to between 90–100°C. Between 3 and 5 drops of glacial acetic acid should be carefully dripped across the sample surface as an initial rinse. This should be followed by a series of soaks at room temperature: acetone, DI water, and IPA soaks for 2 minutes each. The sample should then be gently blown dry with either a compressed air or nitrogen gun. The fuming nitric acid decapsulation process can take 1–2 minutes to complete, and a thorough

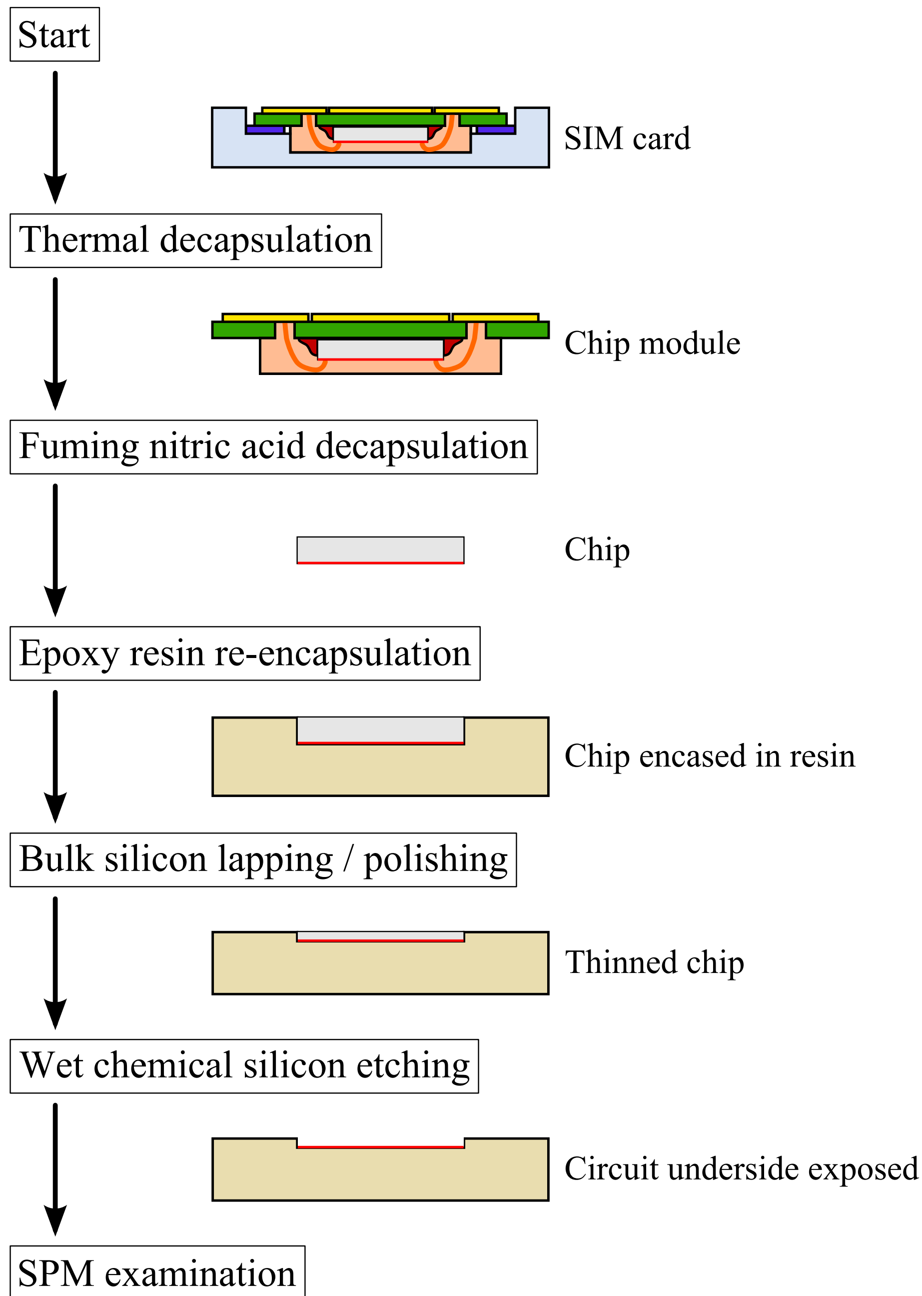


Figure 4.34: Flow diagram outlining the final sample preparation process developed to extract a SIM card microcontroller and process it to expose the underside of the circuit layer.

inspection with an optical microscope should be used to check for completion at regular intervals to avoid over-etching.

At this stage the microprocessor should be thoroughly inspected under an optical microscope. If the die appears undamaged, then mounting it onto a prototyping sample holder and using a wire bonder to connect the contact pads may allow for electrical interrogation of the data. If the chip is notably damaged, or electrical interrogation fails, then the sample should be further processed as below. The thickness of the die should be carefully measured at this point.

Buehler Epo-Thin (2-part) cold-set resin should be mixed in a 5:2 ratio. To aid in bubble removal, the mixture should undergo magnetic stirring and periodic ultrasonic treatments, then left to stand for a few minutes before pouring into the silicone mould. The sample should be set in the bottom of the mould, circuit-side facing up, and the resin carefully poured over it. This cast should be allowed to cure for 10–12 hours at room temperature.

First, the topside of the resin 'puck' should be lapped to remove any meniscus formed in the mould, allowing use of a drop-point micrometer to gauge the depth of remaining silicon. The aim of lapping is to retain just under 40µm of bulk silicon, thin enough for the selective wet chemical etchant to remove without endangering the thin tunnel oxides.

The majority of the bulk silicon should be lapped away very carefully on a rotary water-lubricated grinding wheel using P600 or P800 wet-dry paper. If the sample is set at an angle, this is the stage to carefully correct this issue. When nearing the required depth of silicon remaining, progressively finer grade papers should be used to clean up the surface, smoothing it at each stage while removing silicon at a progressively lower rate. Between each step the sample should be gently washed with water to remove particles that would ruin the finish. After the finest paper has been used, and optical inspection under a microscope reveals no outstanding scratches or pits on the surface, polishing can begin.

Once again, progressively finer particle-size diamond pastes are used on their

designated felt polishing wheels. The sample should be gently washed between each polishing step with mildly soapy water. The polishing process should continue until the finest paste available, at least 1μm grit size.

Following polishing, the sample should be briefly dipped into a buffered oxide etchant solution to remove any native oxide growth, removing every couple of seconds to check for the hydrophobic surface interaction which indicates completion.

The sample should be etched in a solution of TMAH 25wt% Aq. with IPA 10vol% heated in a water bath to a temperature not exceeding 80°C, or a hotplate with magnetic stirrer set no higher than 75°C. At this temperature and concentration the remaining silicon should be etched away in around 2–2.5 hours. The sample must be held away from the etching container sides, always facing upwards and held in the centre of solution. To avoid over-etching, the sample should be checked for the first appearance of circuitry every 15 minutes from 75 mins onwards, and once sighted, every 5 minutes thereafter for completion.

The sample should be gently quenched in warm (≈40°C) DI water for a couple of minutes, then repeated with DI water again at room temperature. The sample should then be gently dipped into acetone, then IPA, and finally very gently blown dry with a nitrogen gas gun, taking care to hold the gun far from the sample. At this stage the sample is prepared for underside AFM examination.

Bibliography

[1] S. Liebert, "Failure analysis from the back side of a die," *Microelectronics Reliability*, vol. 41, no. 8, pp. 1193–1201, 2001.

[2] K. Nohl, "Deep Silicon Analysis," in *HAR*, 2009, p. 27.

[3] M. Kuhn and O. Kömmerling, "Physical Security of Smartcards," *Information Security Technical Report*, vol. 4, no. 2, pp. 28–41, 1999.

[4] W. Byrne, "Three Decapsulation Methods for Epoxy Novalac Type Packages," in *Reliability Physics Symposium, 1980. 18th Annual*, 1980, pp. 107–109.

[5] M. Jacques, "The chemistry of failure analysis," in *17th Annual Reliability Physics Symposium, 1979.* IEEE, April 1979, pp. 197–208.

[6] B. L. Wensink, "Improved technique for decapsulation of epoxy-packaged semiconductor devices and microcircuits," *Microelectronics and Reliability*, vol. 20, no. 4, pp. 537–538, 1980.

[7] F. Beck, *Integrated Circuit Failure Analysis: A Guide to Preparation Techniques.* John Wiley & Sons, 1998.

[8] C. De Nardi, R. Desplats, P. Perdu, F. Beaudoin, and J. Gauffier, "Oxide charge measurements in EEPROM devices," *Microelectronics and Reliability*, vol. 45, no. 9-11, pp. 1514–1519, 2005.

[9] P. Perdu, R. Desplats, and F. Beaudoin, "A review of sample backside preparation techniques for VLSI," *Microelectronics Reliability*, vol. 40, no. 8-10, pp. 1431–1436, Aug. 2000.

[10] S. Murali and N. Srikanth, "Acid Decapsulation of Epoxy Molded IC Packages With Copper Wire Bonds," *IEEE Transactions on Electronics Packaging Manufacturing*, vol. 29, no. 3, pp. 179–183, 2006.

[11] Y. Jiang, X. Song, T. Lam, C. Wu, M. Sun, X. Wang, and X. Li, "MCP Bottom Die Crack Issue during Destructive Analysis," *2007 9th Electronics Packaging Technology Conference*, no. 33, pp. 864–868, Dec. 2007.

[12] T. Guo, P. Nikolaev, A. G. Rinzler, D. Tomanek, D. T. Colbert, and R. E. Smalley, "Self-assembly of tubular fullerenes," *The Journal of Physical Chemistry*, vol. 99, no. 27, pp. 10 694–10 697, 1995.

[13] T. Guo, P. Nikolaev, A. Thess, D. Colbert, and R. Smalley, "Catalytic growth of single-walled manotubes by laser vaporization," *Chemical Physics Letters*, vol. 243, no. 1–2, pp. 49–54, September 1995.

[14] A. Aubert, L. Dantas De Morais, and J.-P. Rebrassé, "Laser decapsulation of plastic packages for failure analysis - Process control and artefact investigations," *Microelectronics Reliability*, vol. 48, no. 8-9, pp. 1144–1148, Aug. 2008.

[15] M. Kruger, J. Krinke, K. Ritter, B. Zierle, and M. Weber, "Laser-assisted decapsulation of plastic-encapsulated devices," *Microelectronics Reliability*, vol. 43, no. 9-11, pp. 1827–1831, 2003.

[16] B. Jones and A. Kenyon, "Retention of data in heat-damaged SIM cards and potential recovery methods." *Forensic Science International*, vol. 177, no. 1, pp. 42–6, May 2008.

[17] 3G Forensics, "DePOT: Epoxy Package Decapsulation - General Useage Instructions," p. 2. [Online]. Available: http://crownhillforensic.com/DePOT_Usage.pdf

[18] C. Wong, "Recent advances in hermetic equivalent flip-chip hybrid IC packaging of microelectronics," *Materials Chemistry and Physics*, vol. 42, no. 1, pp. 25–30, Oct. 1995.

[19] S. Liebert, "Encapsulation of naked dies for bulk silicon etching with TMAH," *Microelectronics Reliability*, vol. 42, no. 12, pp. 1939–1944, Dec. 2002.

[20] Sizes Inc., "Sandpaper (coated abrasives)," August 2004. [Online]. Available: www.sizes.com/tools/sandpaper.htm

[21] Buehler Ltd., "MetaDi diamond paste," 2006. [Online]. Available: www.buehler.com/productinfo/consumables/pdfs/metadi_paste.pdf

[22] S. Takami, Y. Egashira, and H. Komiyama, "Kinetic study on oxidation of si(111) surfaces using h2o," *Japanese Journal of Applied Physics*, vol. 36, pp. 2288–2291, April 1997.

[23] A. Al-Bayati, K. Orrman-Rossiter, J. van den Berg, and D. Armour, "Composition and structure of the native si oxide by high depth resolution medium energy ion scatering," *Surface Science*, vol. 241, no. 1–2, pp. 91–102, 1991.

[24] H. Philipp and E. Taft, "An optical characterization of native oxides and thin thermal oxides on silicon," *Journal of Applied Physics*, vol. 53, no. 7, pp. 5224–, 1982.

[25] L. C. Feldman, P. J. Silverman, J. S. Williams, T. E. Jackman, and I. Stensgaard, "Use of thin si crystals in backscattering-channeling studies of the si-sio$_2$ interface," *Phys. Rev. Lett.*, vol. 41, pp. 1396–1399, Nov 1978.

[26] I. Zubel and M. Kramkowska, "Development of etch hillocks on different Si(hkl) planes in silicon anisotropic etching," *Surface Science*, vol. 602, no. 9, pp. 1712–1721, May 2008.

[27] I. Zubel, "The effect of isopropyl alcohol on etching rate and roughness of (1 0 0) Si surface etched in KOH and TMAH solutions," *Sensors and Actuators A: Physical*, vol. 93, no. 2, pp. 138–147, Sep. 2001.

[28] ——, “The effect of alcohol additives on etching characteristics in KOH solutions,” *Sensors and Actuators A: Physical*, vol. 101, no. 3, pp. 255–261, Oct. 2002.

[29] ——, “Analysis of adsorption of alcohol additives in the process of silicon etching in alkaline solutions,” *Proceedings of SPIE*, vol. 5124, pp. 79–86, 2003.

[30] I. Zubel and M. Kramkowska, “Analysis of Interaction of Surfactant Molecules with Si(hkl) Planes on the Basis of Anisotropic Etching in Alkaline Solutions,” *Acta Physica Polonica A*, vol. 116, pp. S105–S107, 2009.

[31] I. Zubel, “The influence of atomic configuration of (h k l) planes on adsorption processes associated with anisotropic etching of silicon,” *Sensors and Actuators A: Physical*, vol. 94, no. 1-2, pp. 76–86, Oct. 2001.

[32] ——, “Etch rates and morphology of silicon (h k l) surfaces etched in KOH and KOH saturated with isopropanol solutions,” *Sensors and Actuators A: Physical*, vol. 115, no. 2-3, pp. 549–556, Sep. 2004.

[33] ——, “The Model of Etching of (hkl) Planes in Monocrystalline Silicon,” *Journal of The Electrochemical Society*, vol. 150, no. 6, p. C391, 2003.

[34] H. Tanaka, “Effects of small amount of impurities on etching of silicon in aqueous potassium hydroxide solutions,” *Sensors and Actuators A: Physical*, vol. 82, no. 1-3, pp. 270–273, May 2000.

[35] H. Tanaka, D. Cheng, M. Shikida, and K. Sato, “Characterization of anisotropic wet etching properties of single crystal silicon: Effects of ppb-level of Cu and Pb in KOH solution,” *Sensors and Actuators A: Physical*, vol. 128, no. 1, pp. 125–131, Mar. 2006.

[36] V. Korchnoy, “Investigation of Choline Hydroxide for Selective Silicon Etch from a Gate Oxide Failure Analysis Standpoint,” in *28th ISTFA*,

no. 74, 2002, pp. 325–331. [Online]. Available: http://www.engineers.org.il/_Uploads/6850korchnoy.pdf

[37] J. Tsaur, "Investigation of TMAH for front-side bulk micromachining process from manufacturing aspect," *Sensors and Actuators A: Physical*, vol. 92, no. 1-3, pp. 375–383, Aug. 2001.

[38] J. Thong, W. Choi, and C. Chong, "TMAH etching of silicon and the interaction of etching parameters," *Sensors and Actuators A: Physical*, vol. 63, no. 3, pp. 243–249, Dec. 1997.

[39] O. Tabata, R. Asahi, H. Funabashi, K. Shimaoka, and S. Sugiyama, "Anisotropic etching of silicon in TMAH solutions," *Sensors and Actuators A: Physical*, vol. 34, no. 1, pp. 51–57, Jul. 1992.

[40] K. Sundaram, a. Vijayakumar, and G. Subramanian, "Smooth etching of silicon using TMAH and isopropyl alcohol for MEMS applications," *Microelectronic Engineering*, vol. 77, no. 3-4, pp. 230–241, Apr. 2005.

[41] K. Sato, "Anisotropic etching rates of single-crystal silicon for TMAH water solution as a function of crystallographic orientation," *Sensors and Actuators A: Physical*, vol. 73, no. 1-2, pp. 131–137, Mar. 1999.

[42] P. Chen, "The characteristic behavior of TMAH water solution for anisotropic etching on both Silicon substrate and SiO2 layer," *Sensors and Actuators A: Physical*, vol. 93, no. 2, pp. 132–137, Sep. 2001.

[43] M. Shikida, K. Sato, K. Tokoro, and D. Uchikawa, "Differences in anisotropic etching properties of koh and tmah solutions," *Sensors and Actuators A: Physical*, vol. 80, no. 2, pp. 179–188, March 2000.

[44] M. Shikida, "Surface roughness of single-crystal silicon etched by TMAH solution," *Sensors and Actuators A: Physical*, vol. 90, no. 3, pp. 223–231, May 2001.

[45] E. van Veenendaal, K. Sato, M. Shikida, and J. van Suchtelen, "Micromorphology of single crystalline silicon surfaces during anisotropic wet chemical etching in KOH and TMAH," *Sensors and Actuators A: Physical*, vol. 93, no. 3, pp. 219–231, Oct. 2001.

[46] E. van Veenendaal, K. Sato, M. Shikida, A. Nijdam, and J. van Suchtelen, "Micro-morphology of single crystalline silicon surfaces during anisotropic wet chemical etching in KOH: velocity source forests," *Sensors and Actuators A: Physical*, vol. 93, no. 3, pp. 232–242, Oct. 2001.

[47] M. Kramkowska and I. Zubel, "Silicon anisotropic etching in KOH and TMAH with modified surface tension," *Procedia Chemistry*, vol. 1, no. 1, pp. 774–777, Sep. 2009.

Chapter 5

Sample analysis – focused ion beam, Kelvin probe, and sample rewiring

The previous chapter covered the sample preparation process developed to progress from a potentially heat/shock damaged SIM card to a sample consisting of the SIM microcontroller circuit layer removed from the Si substrate, upside-down and ready for AFM examination of the EEPROM/flash memory array. This chapter contains the examination using a focused ion beam sectioning technique to determine the memory architecture expected of SIM card microprocessors. This is followed by a pair of investigations, using the fuming nitric acid decapsulation method developed in the previous chapter, to assess the applicability of a Kelvin probe for initial charge measurement, and to determine the potential data retention characteristics at various elevated temperatures.

5.1 Focused ion beam

The focused ion beam (FIB) is a powerful and versatile tool using a focused beam of ions to mill away and/or deposit material from a sample. It has appli-

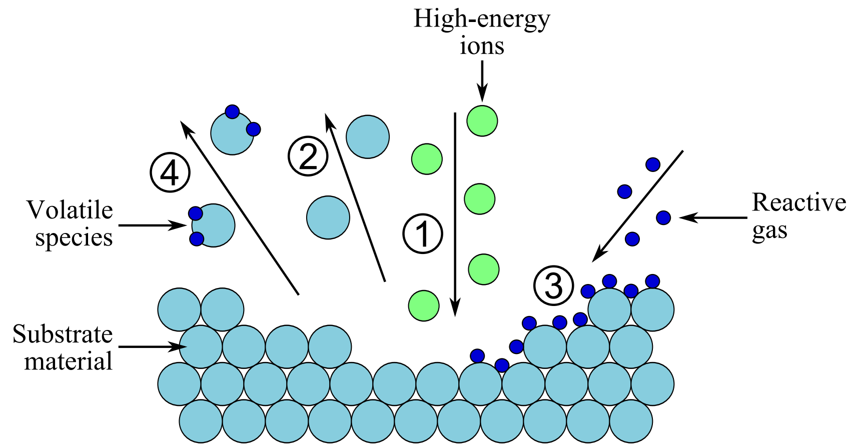


Figure 5.1: Gas-assisted FIB etching process: (1) high-energy ions; (2) sputtered material from substrate; (3) reactive gas adsorption; (4) volatile species formation and evaporation.

cations in a wide variety of fields, predominantly in the semiconductor industry: research, defect analysis, circuit modification and many more areas. Combining the FIB with an SEM system allows highly accurate work to be carried out on the nanometre scale on sensitive devices: an IC could be repaired or modified – cutting connections and depositing new ones; or a sample could be milled away and electropolished to produce a thin TEM sample for analysis.

Ions (often gallium) are large, heavy and slow compared to electrons. Due to their size, ions find it harder to penetrate through the sample surface, exhibiting a far lower penetration depth compared to electrons, but also resulting in a greater degree of ionisation. Gallium ions' greater mass results in a momentum around 370× larger than an electron. This allows material to be removed from the surface of a sample by bombarding it with an ion beam. Magnetic lenses (as used for electrons) are often replaced by more effective electrostatic lenses to handle the increased mass of the ions.

Unlike electron-beam techniques, focused ion beams are destructive to samples. Fig. 5.1 outlines the process involved in gas-assisted FIB etching. By bombarding the surface of a sample with high energy ions (1), substrate material can be sputtered away from the sample (2). This will inherently contaminate

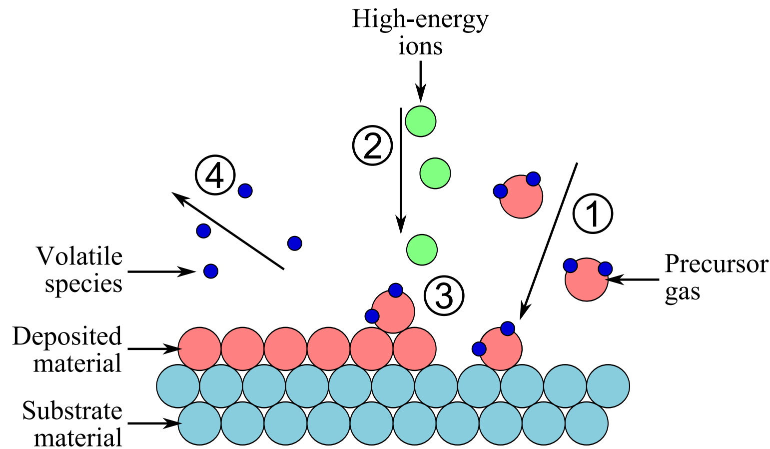


Figure 5.2: Gas-assisted FIB deposition process: (1) precursor gas adsorption; (2) incident ion beam; (3) decomposition of precursor gas, deposits non-volatile component; (4) evaporation of volatile species.

the top few nanometres of the sample with ions. To avoid this radiation damage, an extra layer is often specifically deposited to absorb these ions. Sputtering may be assisted through the interaction of reactive gas molecules with the substrate (3), and the subsequent formation and evaporation of volatile species (4). In addition, the redeposition of sputtered atoms must be considered when using an FIB.

The FIB can also be used to deposit material via a process known as FIB-assisted chemical vapour deposition (CVD), this is outlined in Fig. 5.2. A precursor gas, such as tungsten hexacarbonyl ($W(CO)_6$), is introduced to the vacuum chamber and adsorbs onto the sample surface (1). Careful alignment of the ion beam (2) allows accurate decomposition of the precursor in specific locations (3) into volatile (CO groups) and non-volatile (tungsten) components. The volatile species are then removed (4) while the non-volatile species remain deposited on the surface.

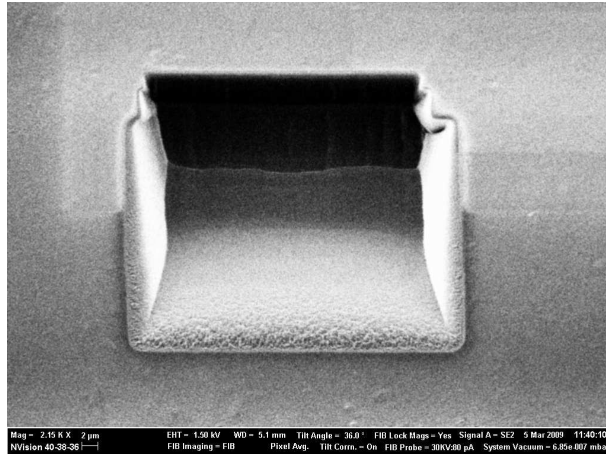


Figure 5.3: FIB trench milled into SIM card microprocessor during cross section imaging.

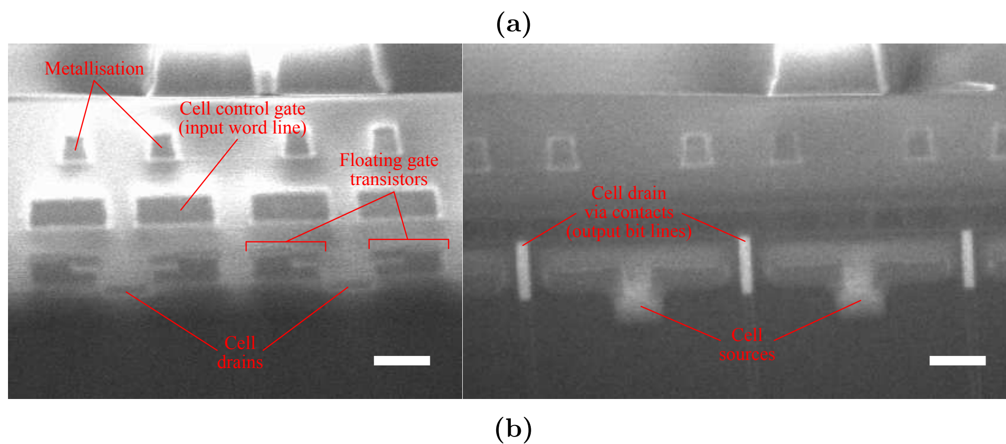
5.1.1 FIB cross-sectional examination of SIM memory array

A SIM card from circa 2003/4, being the oldest by far of the samples obtained, was examined in an FIB-equipped SEM with CrossBeam operation. This was deliberately chosen to allow examination of larger features at this stage of the investigation. CrossBeam refers to the ability to conduct live SEM imaging during milling and polishing steps, the value of which will become apparent. Without knowing what the memory architecture of a specific SIM card would be, the aim of this early work was to identify common features with previously published patents, design schematics, and other SEM examinations by De Nardi *et al.* [1].

FIB milling and cross-sectional imaging was conducted as follows. A wide rectangular trench was roughly milled into the surface of the memory block using a wide, high-power beam. The detector faces into this milled out area, specifically at one of the internal faces in this trench. Further thin layers were repeatedly polished away from this face while under observation using more focused beams, and between each polishing step the SEM took a snapshot of the structure revealed. The material sputtered away is inevitably redeposited onto the other surfaces, mainly building up as a ramp on opposite wall, see Fig. 5.3.

Image removed for
copyright reasons

(a)



(b)

Figure 5.4: Comparison between cross sections of EEPROM and flash structures taken parallel to bit lines. Subfigure **(b)** was examined early in this investigation and shows a NOR flash array (scale bar=1μm), and **(a)** is an EEPROM device from literature [1].

Figure 5.4b shows a cross-sectional SEM image of a central region of a SIM card memory array after FIB milling, with specific parts identified. The remaining EMC on the die surface caused a great deal of charging, introducing a high level of interference in the SEM image. Figure 5.4a shows a cross-section from De Nardi's work [1, Fig. 3]. The results are conclusive, despite differences in the CMOS technology process nodes and device architecture, the cells in both cases show some similar features. The SEM images taken early on in this investigation, Fig. 5.4b, clearly shows a NOR flash architecture consisting of 1T/cell, while Fig. 5.4a shows an EEPROM device showing 2T/cell configuration, with each FGT controlled by an access transistor. Comparing memory array feature sizes, it is estimated that the examined SIM card is manufactured using 180nm CMOS technology, technology first appearing in 1999–2000.

De Nardi's sample was from a 128kbit ST Microelectronics EEPROM memory device with a process node of 0.35µm, circa 1995. This discrepancy in technological development in De Nardi's work from the publication date stems from the ultimate application of their work – astronautics. When choosing technology to send into space for extended durations, because of the high reliability requirements the 'latest and greatest' technology is never used. As such, their publications in 2005/6 examined devices manufactured using the 0.35µm node. To put this into perspective, production of 90nm node devices had begun in 2004, and 65nm node entered production in 2006.

5.2 Applicability of Kelvin probe examination

A study was carried out to assess the applicability of using a Kelvin probe for initial sample examination. Able to measure the surface potential of a whole sample (rather than locally as SKPM would), the technique could prove useful as a means of determining if any data (stored charge) remains after an event such as a house fire. Given the correlation between operating/exposed temper-

ature and charge leakage from the floating gates, using a Kelvin probe to read a sample's surface potential may offer some insight as to whether it is worth continuing sample processing. By measuring samples before and after baking, it was hoped that samples baked at higher temperatures would record a lower surface potential, indicating charge loss.

Obviously there is a caveat, for a real piece of evidence it would be unknown exactly how much data would have been written to the SIM card memory in the first place. However, given that some data is always present, *i.e.* SIM software, this could present a simple but effective technique to numerically determine whether further processing is cost-effective.

The samples had to be programmed, decapsulated, examined with a Kelvin probe, baked at various temperatures, and re-examined with the Kelvin probe. Given the overlap in sample processing required, a second study was conducted using the same prepared and baked samples: to rewire the samples and attempt electronic interrogation to ascertain what temperatures these SIM cards can retain data. Any extracted data would have to be compared to the original data sets to determine the data integrity.

5.2.1 Kelvin probe results

The samples consisted of 3 main types of SIM microprocessor, designated G260, 3GP9, and L254, plus an assorted set of 8 SIM cards of other architectures – totalling 31 samples. The SIM cards were all programmed with sample contacts and SMS data sets, of various sizes depending on each sample's memory capacity, using a USB SIM card reader. Once programmed, the samples were all decapsulated using fuming nitric acid, as described in the previous chapter. Plastic tweezers were safer when handling the microprocessors, but metal tweezers had to be used with fuming nitric acid. As a result, two samples were damaged through mishandling during acid decapsulation.

Each sample's surface potential was initially measured using a KP Technology

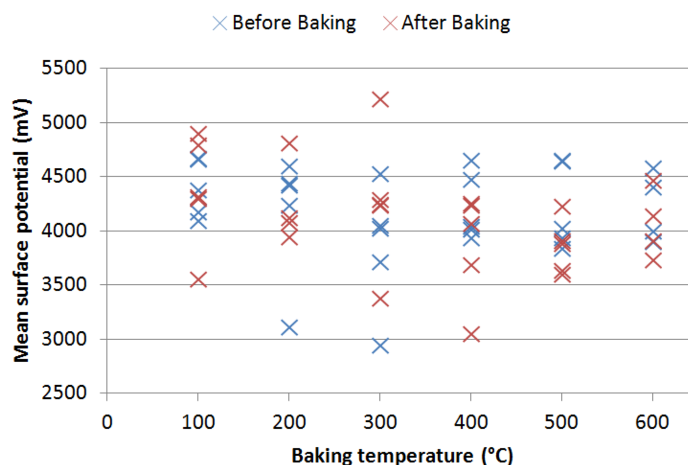


Figure 5.5: Graph of Kelvin probe-measured surface potentials before and after baking at elevated temperatures.

SKP200200 scanning Kelvin probe and a 5mm diameter tip, and then baked in a furnace at specific temperatures for 20 minutes. Before each sample was baked, the furnace temperature was checked with a digital thermometer and K-type thermocouple to ensure accuracy. The chips were placed in a ceramic boat sample holder and slid carefully into the centre of the furnace. An incremental insertion process, with longer ramping at higher temperatures, was conducted for each sample to avoid thermal shock.

This ramping process proved to be very effective, with no samples exhibiting any outward mechanical/thermal shock damage from baking. The samples were re-examined with the Kelvin probe and the results are given in Appendix D. Figure 5.5 shows a graph of the mean Kelvin probe-measured surface potentials before and after baking at temperatures between 100–600°C, while the difference between the surface potential readings, before and after, for each sample at various temperatures is plotted in Fig. 5.6.

As can be seen quite clearly, there is no significant correlation between a sample's surface potential and the temperature it has been exposed to. Some samples increase in potential, while some decrease; and as the temperature that

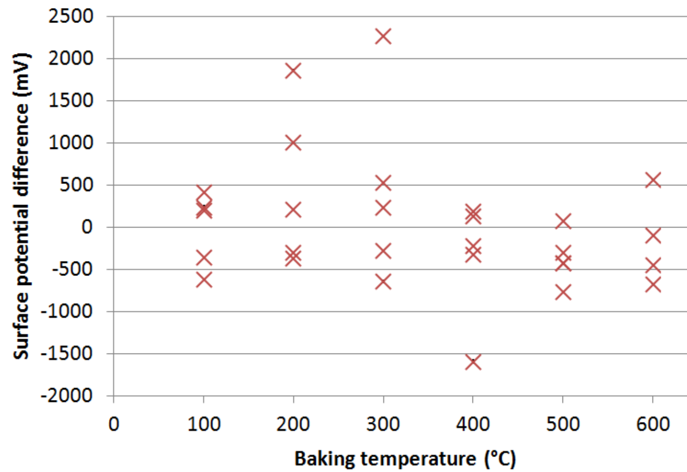


Figure 5.6: Graph of changes in Kelvin probe-measured surface potentials after baking at elevated temperatures.

the samples are exposed to increases, there is no definitive change in the surface potential. Overall, for this application, the Kelvin probe has not proven itself to be a robust enough tool in providing an initial assessment as to the extent of heat-induced charge leakage in a sample. In conclusion, when examining a sample that has undergone unknown thermal exposure, it is best to visually inspect the die for damage after decapsulation processing, and if no serious damage is identified, to attempt rewiring (discussed in more detail below). This conclusion has been included in the final preparation process outlined in Section 4.6 previously.

5.3 Rewiring / data retention

For an approach as extensive and time-consuming as the one proposed in this thesis to be undertaken, other methods of non-destructive interrogation must first be attempted. A feasibility study conducted by Jones and Kenyon [2] heated a dozen decapsulated chips; six to approximately 180°C, five to approximately 450°C, and the final one to approximately 650°C. Some of the samples heated to 180°C suffered mechanical damage and could not be rewired, but those that could allowed extraction of intact, uncompromised data. Given the Arrhenius

relationship discussed in Section 3.4.5, this is to be expected, with data lasting over 12 days at 180°C.

In Jones' work, all but one sample heated to ≈450°C exhibited thermal damage during heating, with the last able to be 'fleetingly operated'. At 650°C the chip fractured, likely due to thermal shock. While the realism of sudden thermal exposure of a microchip to these elevated temperatures is debatable, and would obviously depend highly on a presupposed evidence destruction scenario, the high level of mechanical/thermal shock induced among this sample set is less than helpful.

With this in mind, a second study using the Kelvin probed samples was conducted, aiming to expand upon this initial work by Jones & Kenyon, while also giving a greater context to the full sample preparation process outlined in Section 4.6. These decapsulated SIM card microprocessors that had been heated to various temperatures were attached to prototyping chip carriers, see Fig. 5.7. Using 25µm gold/aluminium wire wedge bonders, the chips' contact pads were wired to the chip carriers' internal connection pads. The chip carriers were attached to custom-designed PCBs, allowing for bespoke fly-wire connections to link specific internally wired contacts to an external PCB header, see Fig. 5.8. These PCBs could then be plugged into one side of a solderless breadboard for further rewiring.

The 'dummy' SIM card consisting of a SIM-sized PCB with corresponding contact pads connected by a flexible flat cable to a custom adapter PCB with headers to the other side of the breadboard is shown in Fig. 5.9a. By finding the pad function combination of the SIM card microprocessors using the breadboard, the dummy SIM could be used to connect a rewired processed SIM to a PC via a USB adapter, see Figs. 5.9b and 5.9c, and read as if it were whole again. This method of electrical interrogation is a recommended step in the process developed in the previous chapter.

In some cases the manual breadboard rewiring was a simple process, involving

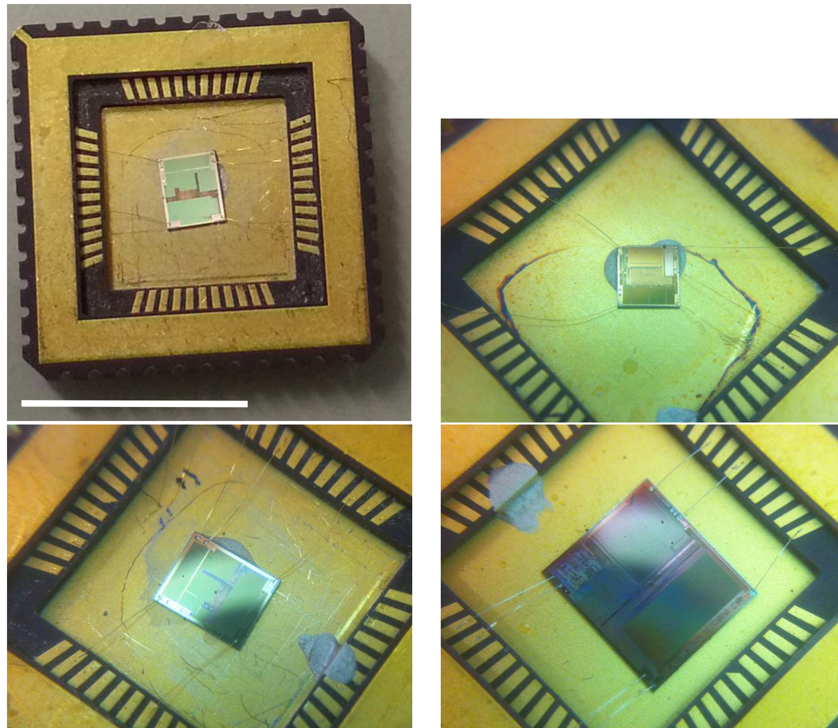


Figure 5.7: Photographs of various SIM card microprocessors wire-bonded to prototyping chip carriers. (scale bar=1cm)

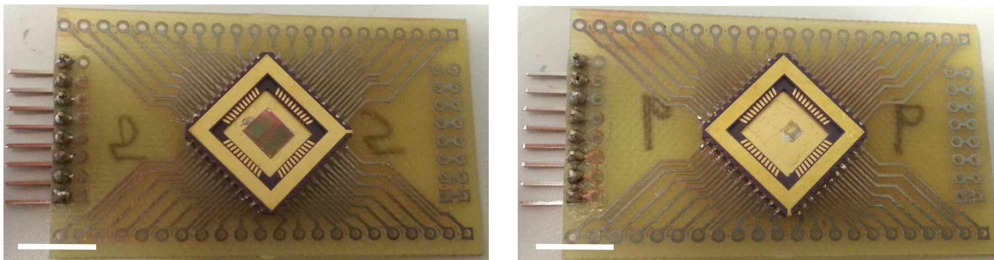
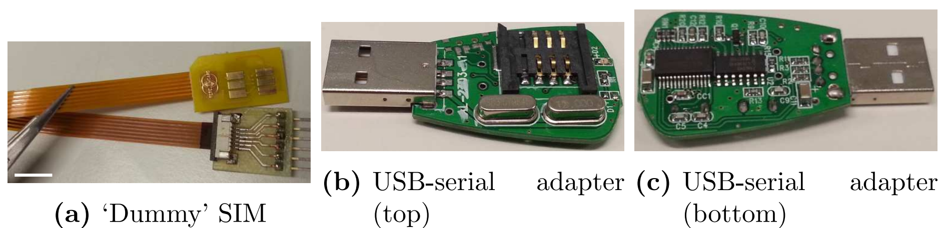


Figure 5.8: Photographs of custom PCB adapters allowing manual through-hole fly-wire connections between specific internal wire-bonded chip carrier ports and header terminals. (scale bars=1cm)



(a) 'Dummy' SIM **(b)** USB-serial adapter (top) **(c)** USB-serial adapter (bottom)

Figure 5.9: Photographs of other adapters used in this study, the 'dummy' SIM card with flexible flat ribbon cable attached to a custom breadboard header (scale bar=1cm), and the USB-serial PC adapter.

optical inspection of the die to identify the pin formation, and thus tracing routes to the board from both ends and making connection. Once the correct permutation of contact pads was identified, the chip could interface with a PC and data could be extracted (if present) and compared to original written data sets for data integrity. However, not all chips had only 5 eligible contact pads – not featuring the now-disused C6 – V_{PP} programming voltage (see Fig. 3.3 in Chapter 3). Extra pads that could not be discounted by optical inspection had to be wired up and tested also. The formula for the possible number of permutations is:

$$\text{Total number of permutations} = \frac{n!}{(n-k)!} \tag{5.1}$$

Where: $k \leq n$; n is the number of possible contact pads; and k is the number of required connections ($n = 5$ for modern SIM cards). Each extra pad made the task of finding the correct permutation exponentially more time-consuming. For $n = 5$, the total number of permutations is 120, for $n = 6$ this rises to 720, and at $n = 7$ this is 2520. Clearly, taking time to conduct careful optical inspection is a time-saver in the long run.

One of the most prominent identifying features is the wear and tear on the contact pads, with newer-looking pads typically being pads dedicated to diagnostic testing or factory firmware programming – these can typically be discounted safely, and only revisited if other permutations fail.

As previously mentioned, the samples consisted of 3 main types, plus 8 additional assorted SIM cards. Two chips that had been damaged during acid decapsulation from mishandling were not included since both had lost a corner (including contact pad) making rewiring impossible. Of the 29 samples baked and rewired, a total of 13 could be successfully read using the method described above. Two samples that could not be read successfully were due to contact pad damage during wire bonding, while one was damaged through mishandling.

Table 5.1: Rewiring success rates after heating a selection of SIM cards to various elevated temperatures.

Temperature (°C)	Success rate
100	60%
200	60%
300	60%
400	40%
500	40%
600	0%

After decapsulation (which requires metal acid-resistant tweezers) the use of vacuum tweezers is recommended, especially for newer SIM cards which can be substantially thinner.

One of the main types (designated L254), and one of the assorted (sample number 30), featured an extensive anti-tampering security mesh across the entire working surface of the chip. None of these samples were able to be read successfully, and it is assumed that the acid decapsulation process used was too aggressive, or mishandling resulted in a break in the thin security mesh. Even one tiny break in these security meshes can render a chip inoperable. The nature of such meshes is to prevent inspection and tampering of the chip, and lock out the device should the mesh be broken at any point. Some meshes work through the use of simple conductive wires – breaking any of the wires will lock the device. More advanced meshes feature resistive and even capacitive sensors to detect advanced tampering, such as using a FIB to cut and reroute mesh wires.

Putting aside the microprocessors with security meshes leaves 19 samples. The larger, and presumably older generation, G260 type chips could be rewired and successfully read up to 300°C. The newer, and smaller, 3GP9-designated samples could be read up to 500°C. The assorted chips could, as expected, be accessed to varying temperatures: one each at 100, 200, 300, and 500°C. The final success rates across the range of temperatures taking into account all encountered causes of device failure are shown in Table 5.1, with final outcomes for individual chips given in Appendix D.

All data extracted from those chips successfully interrogated was intact, identical to the data sets they were initially programmed with. This shows that it is possible to retrieve uncompromised data from some SIM card microprocessors that have experienced elevated temperatures up to 500°C for a period of 20 minutes, and that this is a worthwhile processing step for the recovery of data. Only those unable to be read by rewiring or featuring die damage should be processed further as outlined in the previous chapter. One additional point to take note of is that along with user data corruption, the pre-programmed firmware residing on the microcontroller is also susceptible to heat-induced corruption – should this occur, it may render the device inaccessible despite potentially intact user data.

Bibliography

[1] C. De Nardi, R. Desplats, P. Perdu, F. Beaudoin, and J. Gauffier, "Oxide charge measurements in EEPROM devices," *Microelectronics and Reliability*, vol. 45, no. 9-11, pp. 1514–1519, 2005.

[2] B. Jones and A. Kenyon, "Retention of data in heat-damaged SIM cards and potential recovery methods." *Forensic Science International*, vol. 177, no. 1, pp. 42–6, May 2008.

Chapter 6

Microscopy results

The previous chapter covered the FIB examination of a SIM microcontroller memory array, and the results of investigations into the applicability of using a Kelvin probe for initial sample examination, and temperature-related endurance. This last investigation explored the data retention capabilities of SIM cards exposed to elevated temperatures, and the rewiring process used was similar to that which may take place prior to electrical interrogation of devices with broken bond wires or missing contact pads. This chapter deals with the electrical SPM examination of samples that have been fully processed according to the method created in Chapter 4, *i.e.* samples which have been damaged beyond the point of rewiring.

6.1 Atomic force microscopy

Tapping mode AFM examination was conducted on fully processed samples using a Nanosensors™ SuperSharpSilicon™ non-contact / tapping mode, high resonant frequency AFM probe with topside reflex coating (SSS-NCHR). This high aspect ratio probe is ideal for more accurately measuring the topography of samples featuring tall structures in close proximity to one another, such as the trenches between field oxides in the prepared SIM card samples. The results of a 10μm^2

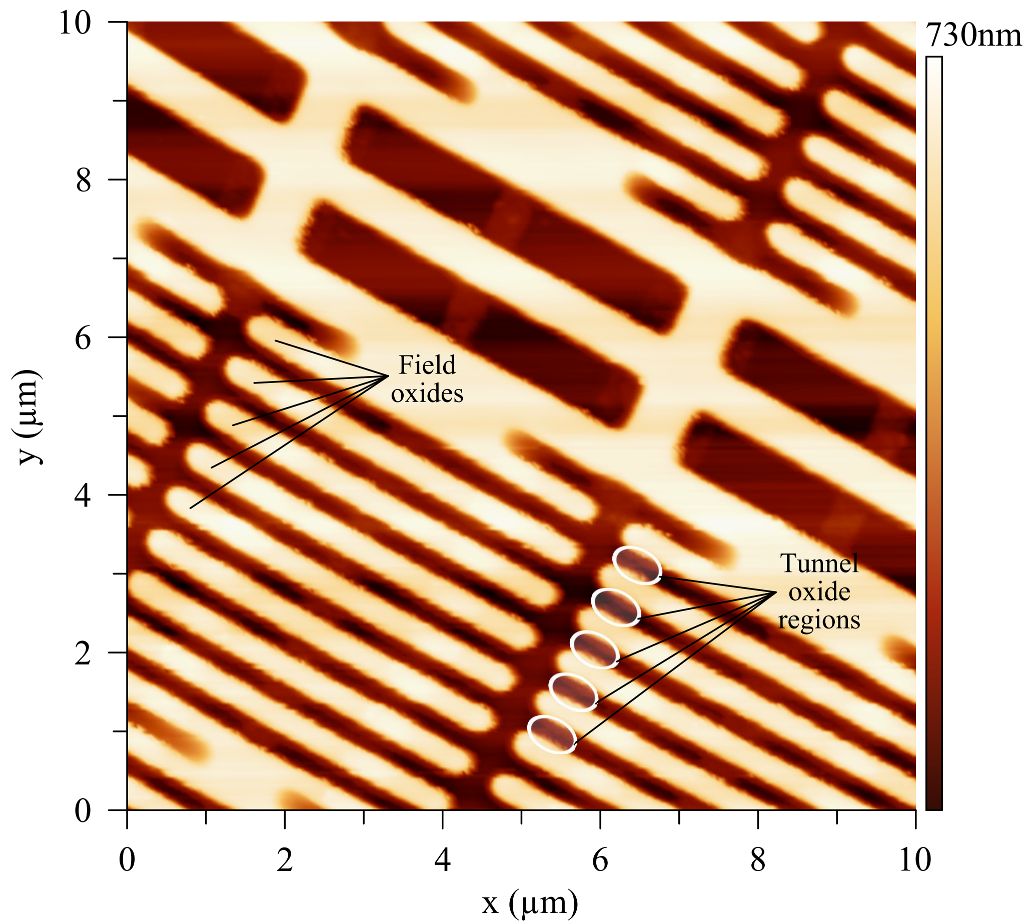
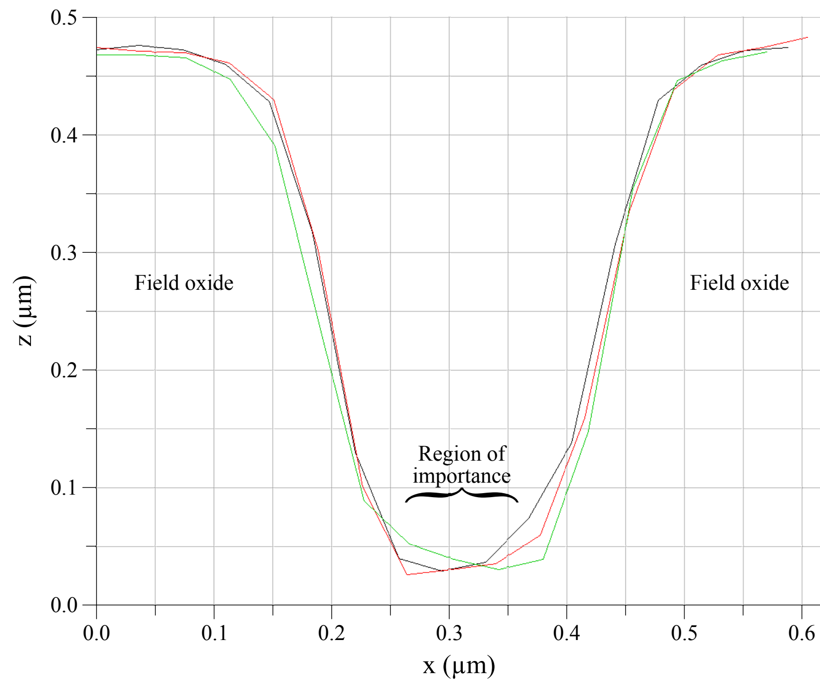
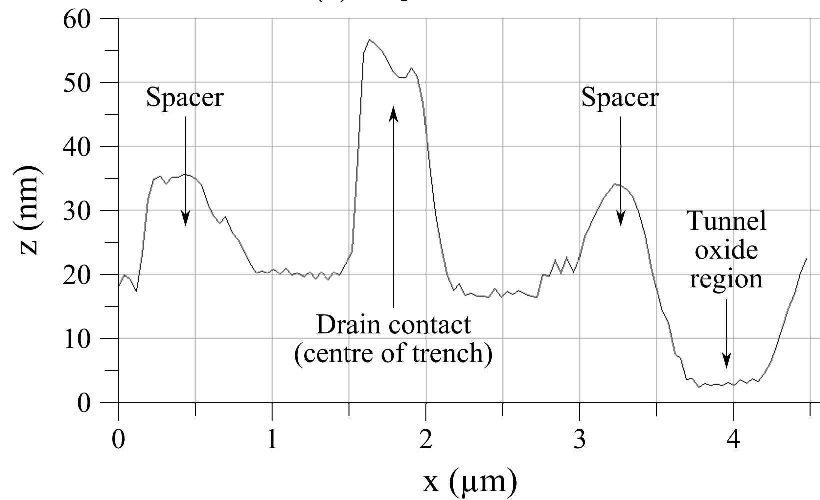


Figure 6.1: Topography scan taken using an ultrasharp AFM tip of an exposed flash memory array – 10µm^2 scan size. Field oxides and approximate locations of regions of interest labelled.

topography scan of the underside of the flash memory array is shown in Fig. 6.1. The field oxide structures running parallel to one another, forming trenches containing the regions of interest are labelled. Also labelled are the approximate locations of some tunnel oxides within these trenches. Linear profiles running perpendicular to the direction of the field oxides have been extracted, and are shown in Fig. 6.2, along with a profile running parallel to adjacent field oxides, along the bottom of a trench. The results have been labelled, illustrating the regions of interest for data extraction.



(a) Perpendicular



(b) Parallel

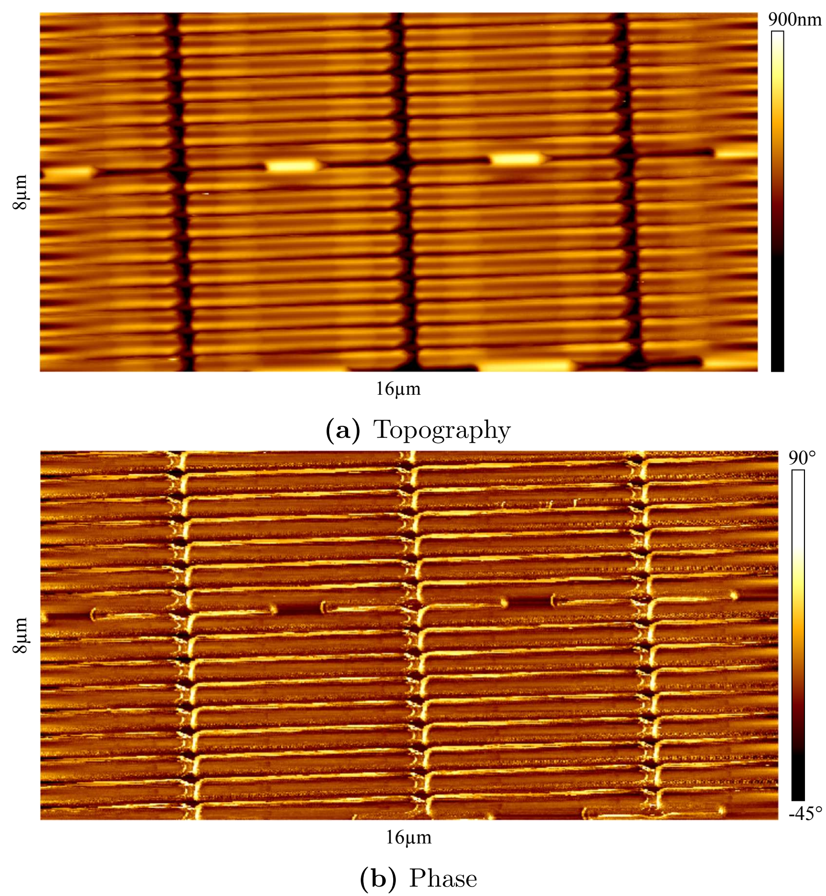
Figure 6.2: Labelled cross-sectional profiles extracted from Fig. 6.1, taken perpendicular to the field oxide lines **(a)**, and parallel to them **(b)** along the bottom of the trench.

6.2 Electric force microscopy

EFM scans were conducted on fully processed samples using specially coated AFM probes – Nanosensors™ PointProbe®-Plus electrostatic force microscopy (PPP-EFM) n^+ silicon cantilevers with Pt/Ir coating on tip and detector side. Figure 6.3 shows two 16×8µm scans showing topography and corresponding lift-mode phase maps. Unfortunately, even at smaller scan sizes than this, the topography of the field oxide regions adjacent to the tunnel oxides introduced edge artefacts, resulting in a high level of interference over the tunnel oxide regions. This interference masked any potentially useful signals relating to the presence of charges within the buried floating gates. The topographical effect of the edges of the field oxides proved to be so extensive that no baseline signal relating to the topography of non-functional regions could be distinguished over the interference.

While EFM has been shown to be sensitive in the detection of trapped charges in other applications, it has been stated previously by De Nardi *et al.* that this mode does not have the capabilities to detect trapped charge within floating gates beneath the tunnel oxide layer of a floating gate transistor memory array [1] (in their scenario, this took the form of an EEPROM device). It has been posited that the reasoning behind this statement is that EFM adopts a deeper probing profile than SKPM. While SKPM measures the shift in CPD required to nullify any electrostatic force encountered by the tip, EFM measures the effect that such a force has on the resonant frequency of the vibrating cantilever.

This presents two scenarios for EFM, both of which result in the tip interacting with a distant electrostatic force. The application of a positive *DC* tip bias acts to attract some of the electrons within the floating gate. Caught between the positively charged tip and the negatively charged proximal region of the floating gate, the surface of the tunnel oxide layer remains neutral, or thereabouts. Thus the electrostatic force being detected arises from charges within the floating gate.



(a) Topography

(b) Phase

Figure 6.3: AFM topography scan **(a)**, and corresponding lift-mode EFM phase scan **(b)** of an 16×8μm section of a SIM card microprocessor flash memory array.

For a negative DC tip bias, the stored charges are repelled, resulting in a close-by neutral region within the floating gate. Few, if any, surface holes appear on the tunnel oxide-substrate interface, and the result is that the electrostatic force felt by the tip is even more distant. The overall result is that for EFM, the invariably biased tip interacts with more distant stored electrons within the floating gate than in SKPM, which interacts with closer capacitively-coupled surface charges. The result is a weaker electrostatic force gradient, which when combined with a sample whose non-planar topography profile introduces extensive artefacts capable of overshadowing the subtle interaction force signal, makes it is easy to theorise why EFM is less compatible that SKPM for this application. Thus far the evidence certainly confirms this result, showing EFM to be unsuitable for the examination of processed flash memory arrays – but as with most things, it should not be completely ruled out until it has been definitively proven to be ineffective.

6.3 Scanning Kelvin probe microscopy

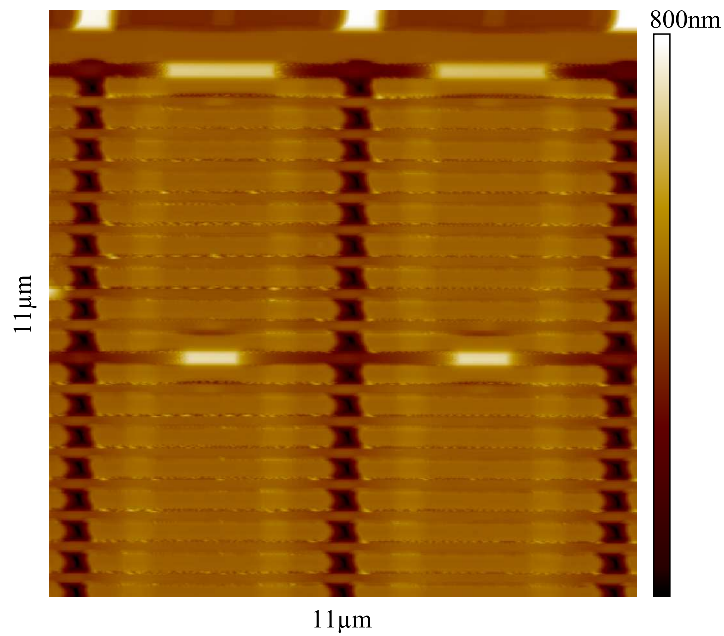
The expected results shown from SKPM imaging of processed SIM chip memory arrays are brighter areas located towards each end of the valleys between field oxide regions. The reason why programmed floating gates, containing injected electrons (negatively charged), would be shown as positively charged bright areas in the potential map is because the tip will not actually be imaging these injected electrons. The presence of electrons within the floating gate repels surface electrons on the opposite (substrate) side of the thin dielectric tunnel oxide layer, resulting in the formation of capacitively-coupled electron holes. Being closer to the tip, these holes have a greater effect on the electrostatic tip-sample interaction during the lift-mode SKPM scans. During SKPM, a feedback loop adjusts the tip bias, equalising the CPD, and balancing any electrostatic forces – effectively removing any tip bias interaction (such as that exhibited in EFM).

The result is that bright spots of positive charge are shown corresponding to the tunnel oxides over negatively charged (programmed) floating gates. In floating gates without additional injected electrons (erased), no capacitively-coupled holes form on the surface, resulting in a region of contrast similar to its surroundings.

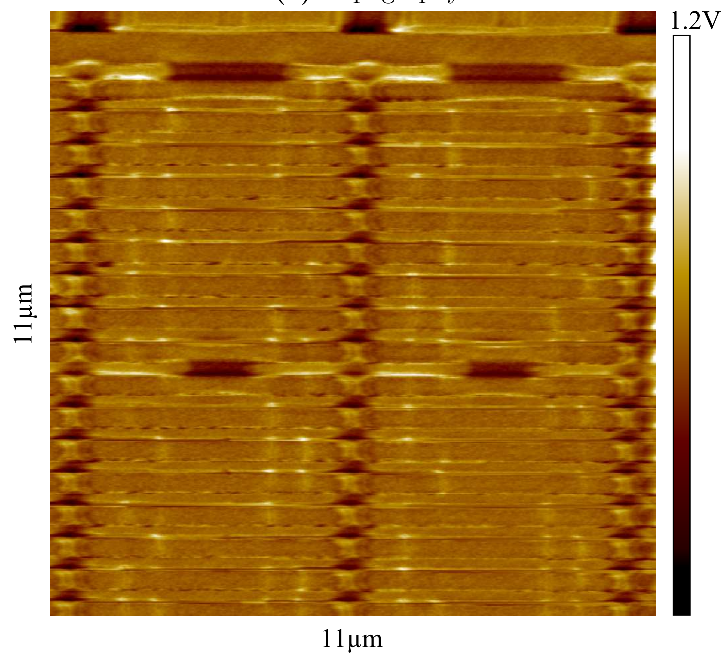
SKPM imaging was conducted using the same Pt/Ir-coated tips as used for EFM. Two 11μm^2 AFM/SKPM scans showing topography and corresponding lift-mode potential maps are shown in Fig. 6.4. The topography, as expected, is still quite visible in the SKPM scan, however, there are a distinct lack of bright spots on the potential scan corresponding to charged gates' tunnel oxide regions. The reasons for this stem from the much smaller feature sizes used on this device over previously published results.

Results by De Nardi *et al.* [1–3] show AFM/SKPM scans from a device, as previously mentioned, manufactured with a 0.35μm process node. An 11μm^2 scan of their device would just fit 6 of the tunnel oxide regions into the scan window, without even managing to fit in their entire field oxide structures; in contrast, the scans shown in Fig. 6.4 can accommodate a section consisting of 64 tunnel oxide regions in 4 complete memory blocks. One advantage this has for any fully developed method dealing with modern devices is that fewer AFM scans (which have a limited maximum scan size of approximately 100μm) are required to examine a flash memory array, and thus fewer errors would be encountered from the process of stitching together multiple scans prior to interpretation. It is estimated that the samples processed in this study had estimated technology nodes of the range 130–65nm, with most around 90nm mark.

The pitch between the field oxides in De Nardi's samples was approximately 3.4μm; the device shown in this study, Fig. 6.4 had a 600nm pitch between field oxides. The estimated side length of their tunnel oxide regions were just over 1.1μm, the devices examined in this study were estimated to be under 100nm. This scaling due to the reduction in technology node, however, means



(a) Topography



(b) Potential

Figure 6.4: AFM topography scan **(a)**, and corresponding lift-mode SKPM potential scan **(b)** of an 11μm^2 section of a SIM card microprocessor flash memory array.

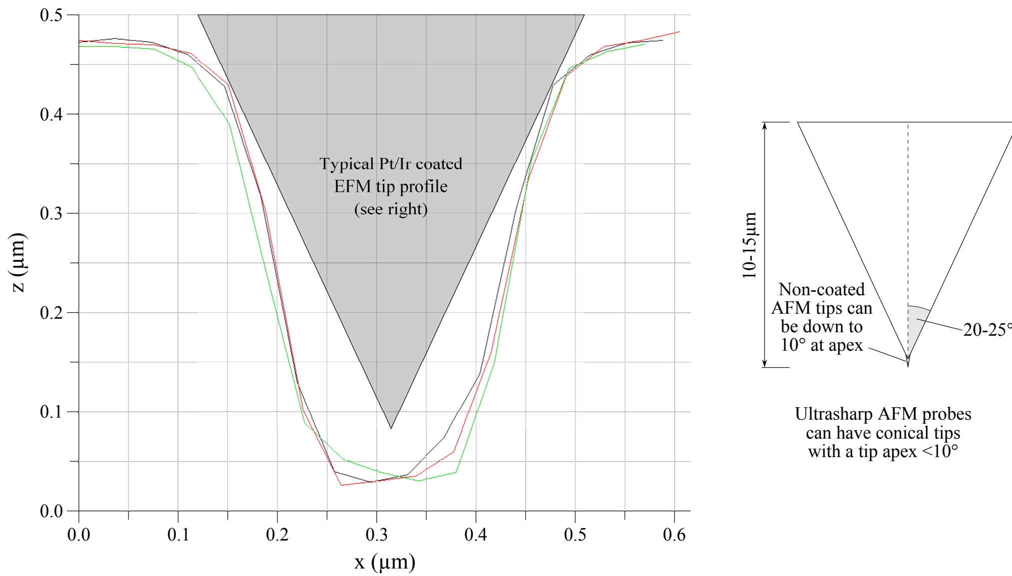
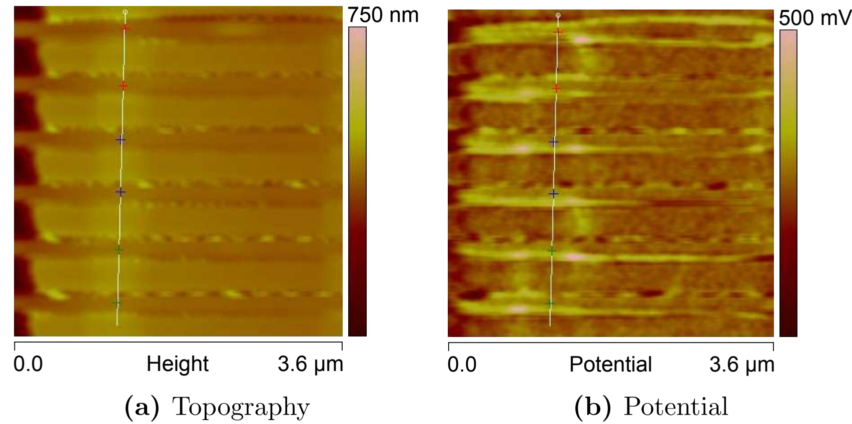


Figure 6.5: Comparison of typical AFM tip profile to previous (ultrasharp) AFM-measured trench profiles.

that topography has a far greater impact on the scan sensitivity. Smaller tunnel oxide regions at the bottom of thinner trenches flanked by closer field oxides will result in a greater degree of interference. This interference would be carried over to the potential scans, disrupting the signal and introducing topography-induced artefacts. Smaller featured memory devices also have a lower ΔV_{T}, *i.e.* fewer electrons are injected and stored in the floating gates during programming operations, making the coulombic field weaker than in previous generations of memory devices, and thus decreasing the signal-to-noise ratio for such samples.

Figure 6.5 shows the extracted cross-sectional profiles (shown previously) taken perpendicular to the field oxides of a flash memory array. A schematic of an AFM tip is shown on the right of the image, and a similar shaped profile to the tip schematic is overlaid in grey on the trench profile. A typical AFM tip, very similar in size and shape to the conductively-coated EFM tips used throughout these investigations, is shown to be unable to image the sample topography at the bottom of the trench. The upper walls of the tip would interact with the field oxide structures, preventing the actual tip from reaching the base of the



(a) Topography

(b) Potential

Figure 6.6: Example of linear profiles taken of topography perpendicular to the field oxides **(a)**, and marked to locate the regions of interest (in this case a control region). This allows the user to swap the data channel to the SKPM surface potential data **(b)**, and accurate measurements to be recorded of the potential at these points.

trench. In multiple-pass electrical SPM techniques this would act to disrupt the initial topography image. This variation would result in the secondary lift-mode scan following a different path from the actual topography of the trench regions. The secondary lift-mode scan, following this erroneous path at a raised height, would now not be following a path of constant van der Waals forces, and thus the tip-sample electrostatic force interaction signal would be disrupted dramatically, potentially allowing the van der Waals forces to once again become the dominant tip-sample interactions.

Despite producing a shadow of large topographical features, *e.g.* field oxides, the potential map will not allow for accurate positioning over small features. In practice, cross-sectional profiles were taken through multiple features in the initial AFM topography scan, and regions of interest were marked accordingly. With these regions marked, the data channel could be swapped to show the lift-mode SKPM surface potential data, and the markers would pinpoint the regions of interest. The SKPM-measured surface potential at the tunnel oxide regions of the FGTs was recorded, along with a nearby control region (as shown in Fig. 6.6). Both profiles can also be plotted together and directly compared if

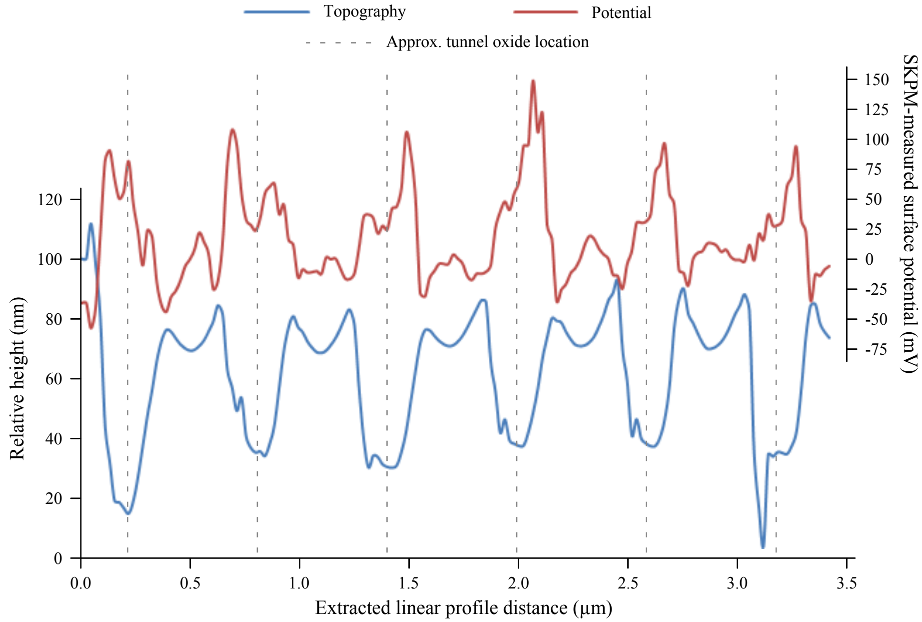


Figure 6.7: Comparison of extracted topography and potential cross-sectional profiles taken perpendicular to the field oxides.

so desired; an example of such a comparison is shown in Fig. 6.7.

The average 'baseline' potential was calculated for each bit line by averaging the potential of nearby non-functional areas along that line. This figure was then subtracted from each specific measurement (both tunnel oxide and control regions) for that bit line. This had the effect of normalising the readings, allowing direct comparison of surface potential measurements.

Despite earlier concerns, in Fig. 6.4 the tip did appear to image the base of the trenches successfully at the mid-points between adjacent field oxides. Although the surface potential map lacks the obvious brightly contrasted tunnel oxide regions, such as those shown by De Nardi *et al.* [1–3], a clear difference can be seen in the statistical distributions of the surface potentials measured over the tunnel oxide regions, below which lie the isolated floating gates (potentially) containing injected electrons. The distributions of the floating gate surface potentials form a double-peaked histogram that can be fitted with twin-Gaussian distributions corresponding to the two different (logical) transistor states, see Fig. 6.8. Measurements taken over floating gates containing injected electrons

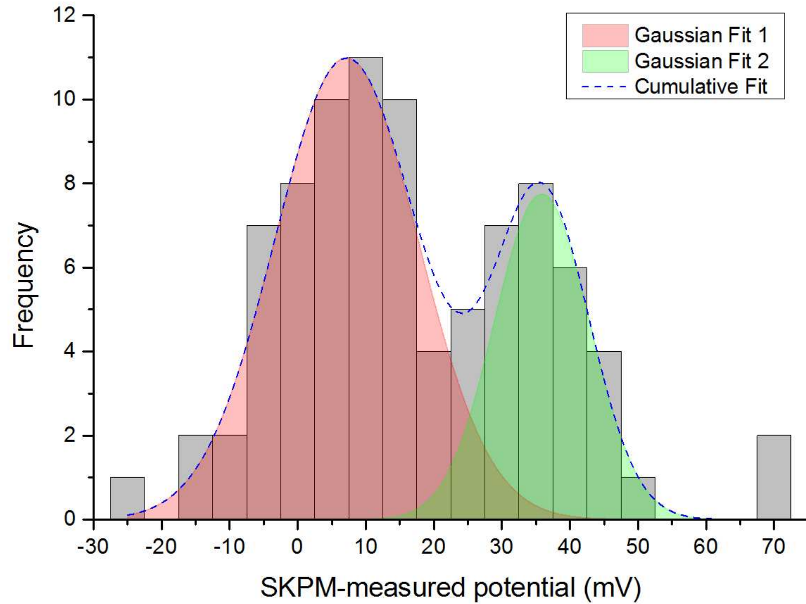


Figure 6.8: Twin-peaked histogram of normalised surface potential measurements over tunnel oxide regions, fitted with two Gaussian distributions to better show different logical states of the transistors: the lower potential red curve corresponds to a logical 1 (erased) state, and higher potential green to a logical 0 (programmed) state.

Table 6.1: Gaussian fit parameters for the three distributions.

Distribution	Mean Potential (mV)	St. Dev.
Control region	14	15
FGT, logical 1	7	10
FGT, logical 0	36	7

form a more positively-shifted distribution compared to the distribution formed by erased floating gate measurements.

In contrast, the normalised surface potential measurements of a nearby 'control' region for each bit line can be similarly plotted in a histogram and fitted with a Gaussian curve. The results shown in Fig. 6.9 show what one would expect to see of such a region – a single-peaked Gaussian distribution. For comparison, the individual Gaussian distribution parameters for all three Gaussian fits of the data is shown in Table 6.1.

Comparing the mean potential of the erased (logical 1) FGT and control

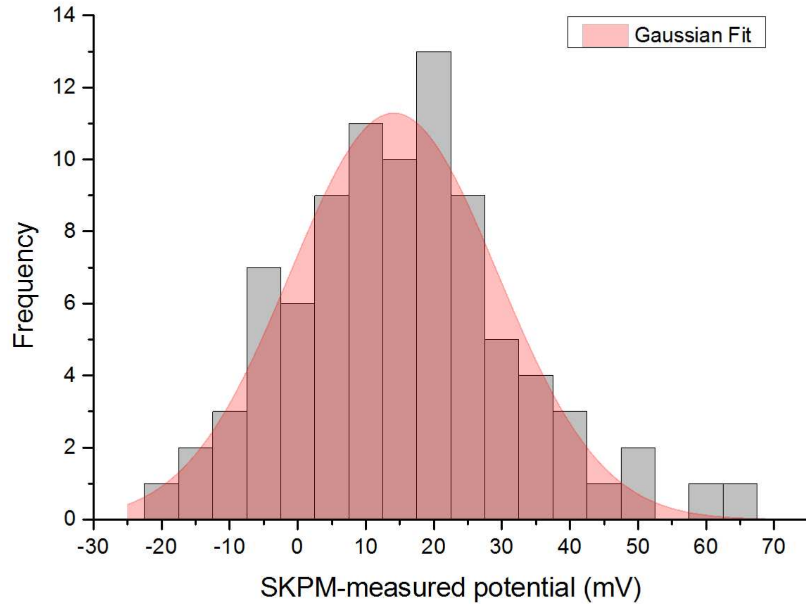


Figure 6.9: Single-peaked histogram of normalised surface potential measurements over control regions near tunnel oxides, fitted with a Gaussian distribution.

region distributions shows the mean of the Gaussian fit to be slightly lower (≈7mV) for the cells containing erased floating gates. From this it is possible to conclude that the floating gate may be slightly over-erased during operation – a common result after erasing whereby not only are stored electrons removed from the floating gate, but some intrinsic electrons are also forced out by the operation, leading to the formation of holes. However, this could be presumptive, after all the two regions consist of very different structures – one an electrically isolated floating gate able to store injected electrons, and the other an arbitrarily chosen control region nearby. There is also the difference in material to consider, and such effects could explain the small difference between them.

From the twin-Gaussian fit in Fig. 6.8 it is possible to estimate a distribution logic margin window between programmed and erased states of 29±5mV. This is unlikely to be the actual logic margin window size as this is just the SKPM-measured mean surface potential difference between states. To illustrate this, consider previous results published by De Nardi *et al.* in 2005. Their result

of 100mV difference between logical states is highly improbable given that the actual logic margin window for EEPROM/flash memory would have been over ten times this figure – of the order of volts.

Unlike De Nardi's results, which were able to directly compare adjacent transistors, the 29mV figure derived from examining flash memory arrays represents the *average* SKPM-measured surface potential difference between possible transistor states. Without an ideal sample situation and accurate calibration within a single scan, it is only relatively quantifiable and cannot be directly compared numerically to the device's actual electrical logical margin window. On older samples this could produce distinctive bright spots indicative of the presence of stored charge beneath the thin tunnel oxide layer; on more modern samples, scaling appears to have had the effect of reducing these obvious stored charge states to mere statistical distributions, which are of less use to a forensic examiner seeking definitive and conclusive data extraction.

With greater development this technique could potentially improve upon the results shown above. Increasing the sensitivity of SKPM measurements and decreasing the level of topography-induced interference should have the effect of shifting the programmed (logical 0) distribution further from the erased (logical 1) distribution. This would allow greater discrimination between logical states, perhaps resulting in similar 'bright spot' results shown to be possible when imaging older technology EEPROM devices.

Bibliography

[1] C. De Nardi, R. Desplats, P. Perdu, F. Beaudoin, and J. Gauffier, "Oxide charge measurements in EEPROM devices," *Microelectronics and Reliability*, vol. 45, no. 9-11, pp. 1514–1519, 2005.

[2] ——, "EEPROM Failure Analysis Methodology - Can Programmed Charges Be Measured Directly by Electrical Techniques of Scanning Probe Microscopy?" in *31st ISTFA 2005*. San Jose, CA: ASM International, 2005, pp. 256–261.

[3] C. De Nardi, R. Desplats, P. Perdu, J. Gauffier, and C. Guerin, "Descrambling and data reading techniques for flash-EEPROM memories. Application to smart cards," *Microelectronics and Reliability*, vol. 46, no. 9-11, pp. 1569–1574, 2006.

Chapter 7

Quartz grain analysis

The AFM can be used in a number of ways in the forensic analysis of evidence. As well as use in electronic forensic applications, as previously shown, the capability of AFM to provide high-quality topographical data in a quantifiable form has application to many other branches of forensics. The work discussed in this chapter is an application of AFM to one such field – the quantitative identification and distinction of the provenance of quartz sand grains.

Much of the information in this chapter has been published in a similar format in *Investigation of quartz grain surface textures by atomic force microscopy for forensic analysis* [1] published in Forensic Science International. New data from the original sample sets plus an entirely new sample set has been included. This data has been analysed accordingly in the final discriminant analysis conducted later in the chapter.

7.1 Introduction

The analysis and interpretation of soil/sediment samples can provide indications of their provenance [2, 3] and enable exclusionary comparisons to be made between samples pertinent to a forensic investigation. There are many analytical techniques (physical, chemical and biological) that can be utilised to compare

such samples, and the analysis of the morphological textures on quartz grain surfaces is one technique that has been demonstrated to be able to provide highly discriminatory results in forensic cases [4, 5].

Quartz grains are highly resistant, but the surfaces of each grain record its history and the environments to which is has been exposed. It is possible to identify features that are indicative of its original parent rock, the transportation mechanisms it has been exposed to, and the processes that have been operating on it in a new deposited environment. In forensic cases, quartz grain surface textures have also been shown to be able to survive anthropogenic factors such as vehicle fires [6]. In forensic casework, the most common use of quartz grain surface texture analysis is in the comparison of soil/sediment samples from different sources (crime scene, suspect, alibi site) to discriminate in an exclusionary manner between quartz derived from different samples [7]. On occasion it is used to derive the provenance of an unknown location in so called 'seek and find' investigations [7].

Whilst the use of quartz grain surface texture analysis in forensic casework studies has been well-documented, it is a technique that is onerous in terms of the operator time required, relies on an experienced expert operator, and the SEM images produced are not typically quantifiable outputs. To this end a number of studies have been undertaken to assess the potential for alleviating some of these difficulties with specific reference to the use of this technique within a forensic context [8, 9].

As discussed at length in Chapter 2, the AFM is a form of scanning probe microscope utilising an ultrasharp tip, micro-fabricated on a flexible cantilever to image a sample. When in close proximity with the sample, the deflection of the cantilever, due to interaction with intermolecular forces between the tip and the sample surface, is detected by reflecting a laser beam off the top of the cantilever onto a position-sensitive photodiode. Given a known deflection and spring constant for the cantilever, the interaction force between tip and sample

can be obtained using Hooke's law. Feedback circuitry controls the tip-sample distance to remain small, but also to avoid damaging the tip/sample. Positioning of the tip is precisely controlled with piezoelectric elements in the x, y, and z axes, and during analysis the tip is scanned across the sample surface in a raster scan pattern, gradually building up an image of the surface topography. The AFM can be operated in different scanning modes according to the dominant intermolecular force regime. AFM is, therefore, a powerful technique in the study of surface analysis allowing nanometre-scale resolution and quantitative measurement of surface features. Generated data can readily be analysed using simple statistical methods to provide quantified figures of merit to aid in sample identification and discrimination.

Due to the high resolution and quantifiable results, the AFM has been applied to a range of disciplines over the years, including a variety of fields within the forensic sciences. It has, for example, been utilised in the forensic examination of the surface of hairs [10, 11], textile fibres [12], and line crossings [13]; age determination of blood stains [14, 15]; investigation of fingerprints on metallic surfaces [16]; and spectroscopic force microscopy of commercial pressure-sensitive adhesive tapes [17].

Given the recent criticisms levied against many forensic science disciplines [18, 19] there is arguably now a drive to provide independent corroboration of forensic analysis to enable such evidence to have sufficient weight in a courtroom. Analysis of quartz grains using the AFM has the potential to provide a quantitative analysis tool that could provide independent and corroborative analysis to complement the morphological surface texture typing of quartz grains achieved by SEM in forensic cases. The aim of this study is therefore, to establish the utility of AFM as a tool for the quantitative examination of quartz grain surface textures specifically for forensic exclusionary comparisons.

7.2 Materials and methods

7.2.1 Instrumentation

All AFM measurements were performed on a Veeco Dimension 3100V with a hybrid x-y-z scan head and using NanoScope software version 7.30. The cantilevers used were: Nanosensors™ PointProbe®-Plus non-contact / tapping mode (PPP-NCLR) n^+ silicon cantilevers; Nanosensors™ PointProbe®-Plus diamond-coated non-contact / tapping mode (DT-NCLR) n^+ silicon probes; and Advanced Diamond Technologies© NaDiaProbe® all-diamond probes. All images were taken in tapping (intermittent-contact) mode unless stated. All imaging was conducted under ambient conditions.

7.2.2 Samples

Three sets of samples were chosen for the original analysis; see Figs. 7.1 to 7.3 for SEM images of sample grains. The first sample was crushed pure Brazilian quartz and exhibited very characteristic clean, fresh faces and sharp edges. Sample 2 was derived from a body deposition site from a concluded murder investigation in the UK. The grains within this sample were distinctive with the presence of euhedral crystal growths and naturally formed sharp edges caused by diagenetic processes. Sample 3 was also derived from a different concluded forensic investigation collected from an alibi site. The sample was characterised by semi-round, aeolian grains with characteristic upturned plates and conchoidal fractures.

Twenty-four grains from the three sample sets were examined under AFM, each grain imaged from multiple angles. The grains were pressed into pressure-sensitive adhesive putty to provide adequate support during the scanning process; this allowed for retrieval of the grains afterwards, or realignment if necessary during analysis.

A fourth sample set was later obtained, originating from a crime scene at a

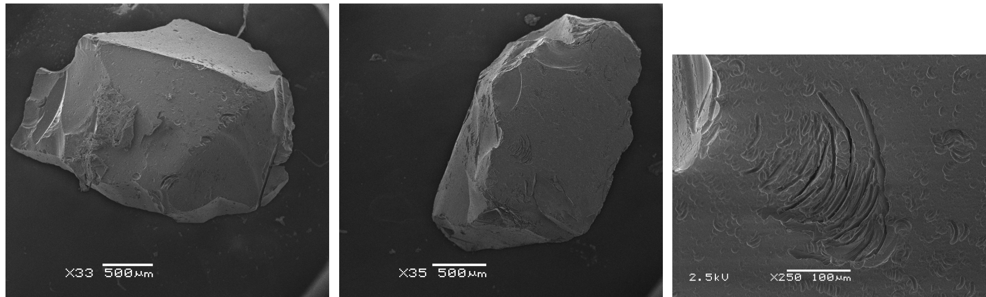


Figure 7.1: SEM images of grains from sample set 1 – Mechanically crushed pure quartz. Sharp edges and clean faces with minimal mechanical surface textures.

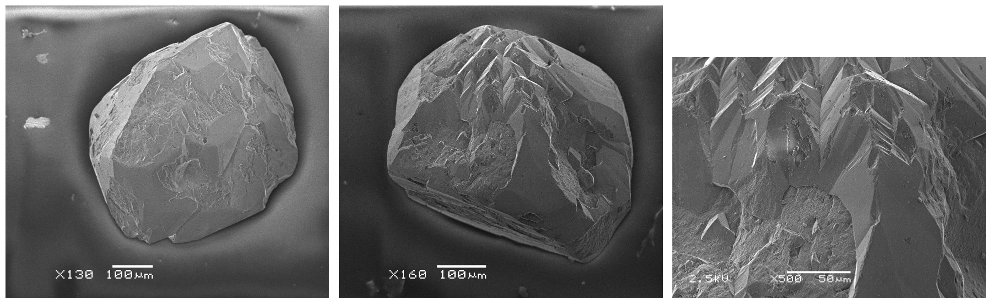


Figure 7.2: SEM images of grains from sample set 2 – Diagenetic quartz with euhedral crystal growths formed naturally by deposition under high pressure. Minimal transportation of the grains is evident from lack of edge abrasion.

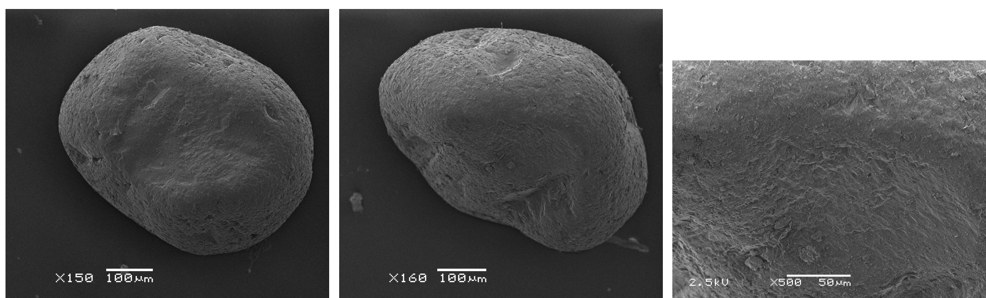


Figure 7.3: SEM images of grains from sample set 3 – Aeolian quartz grains. Semi-round grains with characteristic upturned plates on the surfaces in addition to conchoidal fractures.

beach in Skegness – thus subjected to both aeolian and aqueous coastal erosion. Data derived from AFM scans of this sample were added to the original three data sets, and an overall analysis was conducted.

7.3 Results

7.3.1 Topographical AFM scans

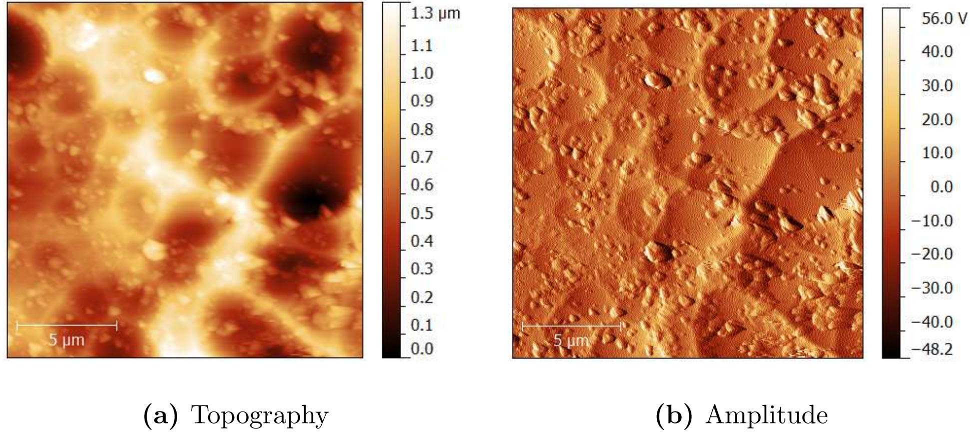
Figure 7.4 shows AFM topography and deflection amplitude signals obtained from a grain in sample set 1 showing a surface riddled with impact craters formed from mechanical crushing. Note that topographic data are derived from tip deflection data. Offset from this scan, Fig. 7.5 shows the interface between two very different surface textures on sample 1. The right-hand side of the image shows the cratered surface exhibited in Fig. 7.4. In contrast, the left shows a recently cleaved crystal face with a highly-ordered striated structure. A 3D reconstruction to better visualise the surface is given in Fig. 7.5c.

The grains in sample 2 were formed naturally by deposition under high pressure and exhibit sharp edges, as can be seen in Fig. 7.6. Scattered across the sample surface are euhedral crystal growths characteristic of this type of grain formation. Figure 7.7 is a scan of a relatively flat surface region with a large number of these crystal growths visible, particularly in the accompanying amplitude scan.

Aeolian grains such as those in sample 3 are characterised by significantly rounder edges and a structure consisting of upturned plates overlapping one another, this can be seen in Figs. 7.8 and 7.9.

7.3.2 Statistical analysis

Once the topography scan of the grain surface has been taken it is relatively simple to compute basic statistical functions on the data, with a range of standard calculations often provided by the scanning software.



(a) Topography **(b)** Amplitude

Figure 7.4: AFM topography scan **(a)** and accompanying amplitude map **(b)** obtained from a grain in sample set 1 exhibiting impact craters resulting from mechanical crushing. Scan size 17.5µm^2.

Table 7.1: Computed roughness values for the two different types of grain surface shown in Fig. 7.5.

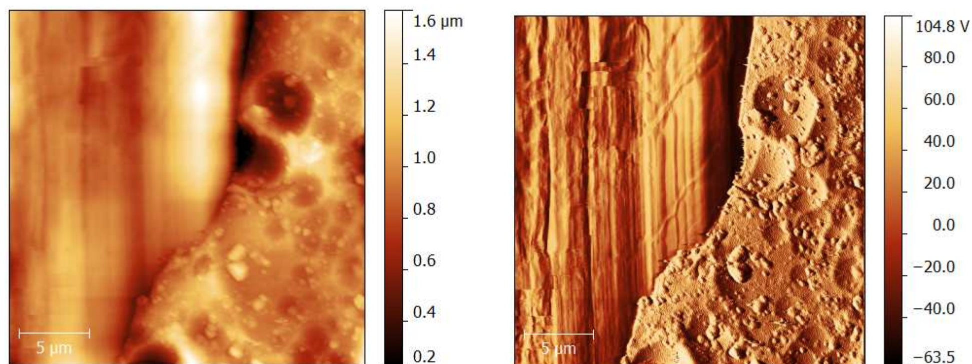
Area	S_a	S_q
Cratered region	0.127	0.177
Crystalline region	0.149	0.193

For Fig. 7.5, visual inspection would indicate the heavily-cratered surface on the right to be rougher than the adjacent freshly-sheared crystal face on the left. However, after taking the arithmetic average (Eq. (7.1)) and root-mean-square (Eq. (7.2)) surface roughnesses of the topography scan, the cleaved crystal face is shown to be numerically rougher – values are given in Table 7.1.

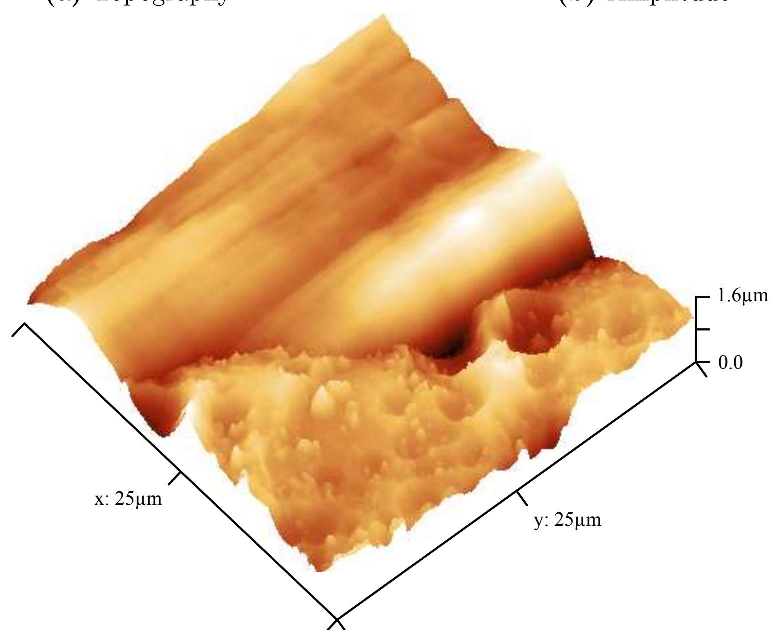
Arithmetic average surface roughness
$$S_\mathrm{a} = \frac{1}{mn}\sum_{j=1}^{m}\sum_{i=1}^{n}|z(x_i, y_j)| \tag{7.1}$$

Root-mean-squared surface roughness
$$S_\mathrm{q} = \sqrt{\frac{1}{mn}\sum_{j=1}^{m}\sum_{i=1}^{n}z^2(x_i, y_j)} \tag{7.2}$$

Where: m is the number of points per line, n is the number of lines, and z is the height at (x, y) point. It is also possible to provide measures of surface skewness (Eq. (7.3)) and kurtosis (Eq. (7.4)) of the grain surfaces.

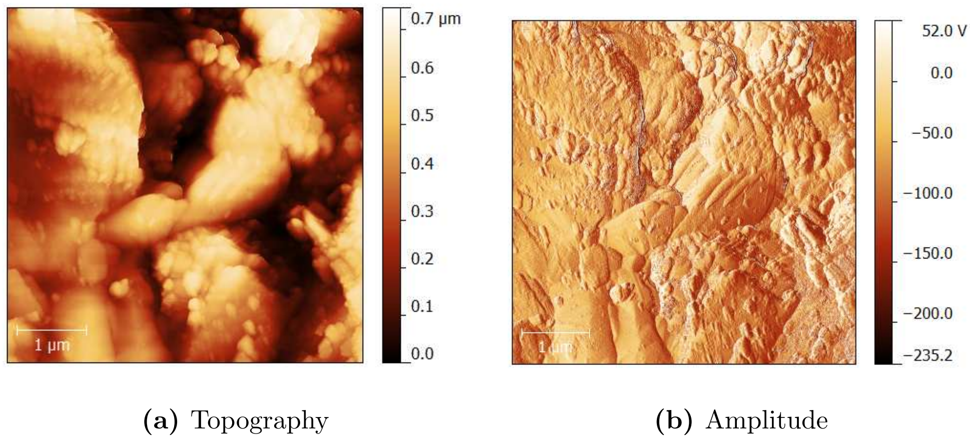


(a) Topography **(b)** Amplitude



(c) 3D topography reconstruction

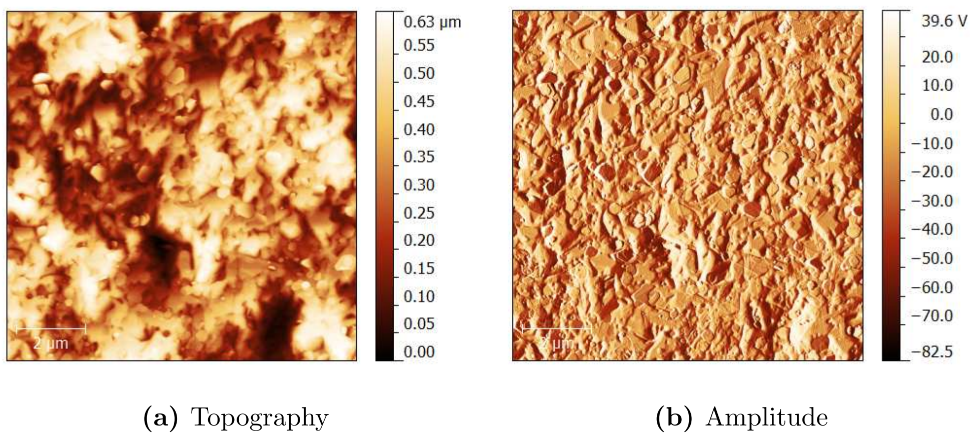
Figure 7.5: AFM topography scan **(a)**, accompanying amplitude map **(b)**, and false colour 3D elevation map **(c)** offset from Fig. 7.4. Clearly visible is the interface between two different surface textures. Scan size 25µm^2.



(a) Topography

(b) Amplitude

Figure 7.6: AFM topography scan **(a)** and amplitude map **(b)** obtained from a grain in sample set 2 exhibiting sharp edges formed under high pressure. Scan size 5µm^2.



(a) Topography

(b) Amplitude

Figure 7.7: AFM topography scan **(a)** and amplitude map **(b)** obtained from a grain in sample set 2 exhibiting a large number of euhedral crystal growths. Scan size 10µm^2.

Figure 7.8: Topography map obtained from a grain in sample set 3 showing upturned plates formed by wind abrasion. Scan size 30µm × 50µm.

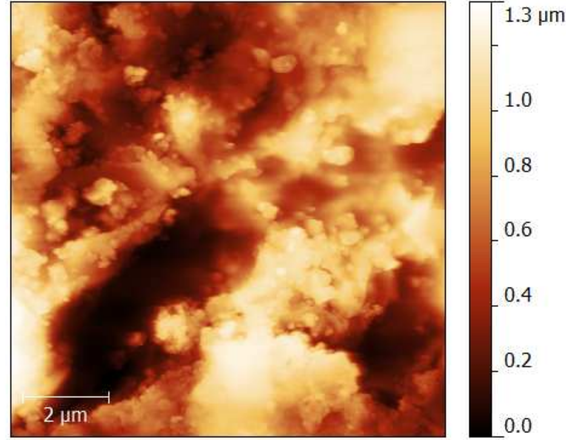


Figure 7.9: Topography map obtained from a grain in sample set 3 showing similar upturned plates and a rounding of edges. Scan size 10µm^2.

$$\text{Surface skewness} \quad S_{\text{sk}} = \frac{1}{mnS_{\text{q}}^3} \sum_{j=1}^{m} \sum_{i=1}^{n} z^3(x_i, y_j) \tag{7.3}$$

$$\text{Surface kurtosis} \quad S_{\text{ku}} = \frac{1}{mnS_{\text{q}}^4} \sum_{j=1}^{m} \sum_{i=1}^{n} z^4(x_i, y_j) \tag{7.4}$$

Further analytical tools, such as power spectral density, can be used to differentiate between two surfaces with equal average feature heights. This allows differentiation according to feature distribution in addition to roughness.

After taking a topographical scan of the surface, it is possible to construct a histogram of the heights of the data points. Applying a Gaussian low-pass filter to smooth the results, the resulting height distribution histogram gives the relative frequency of specific height data obtained from the surface and this can aid in the comparison of different surfaces. For a surface of randomly varying topography, these height distribution histograms would typically take the form of a skewed Gaussian distribution. Deviations from the typical Gaussian shape gives information about the distribution of the surface topography, *e.g.* a double-peaked distribution will signify two distinct levels of surface topography, which

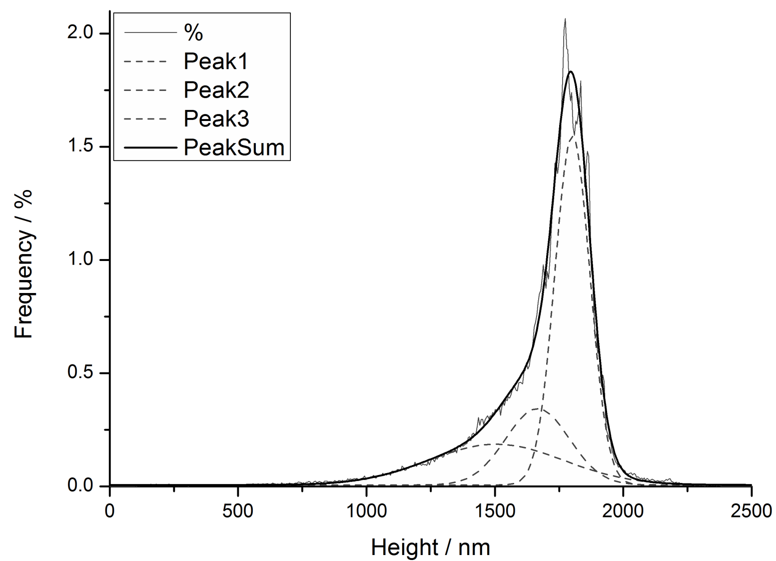
can be further analysed by calculating the difference in the peak mean values – giving a value for the average difference in heights between both regions.

Although this does not substitute for direct cross-sectioning of the sample image to analyse feature sizes, it provides a tool to compare surfaces in a statistical manner using measures such as skewness and kurtosis. Skewness is a measure of the asymmetry of the surface height data distribution, with surface protrusions above the average level resulting in positive values, and surface depressions resulting in negative skewness values. On the other hand, kurtosis describes the 'peakedness' of a surface, *i.e.* a figure describing whether the surface contains extreme height features, with positive kurtosis indicating a few very high peaks or very low troughs, and negative kurtosis indicating moderate deviations in height. Canonical discriminant function analysis was conducted on statistical results from the height distribution and the topographic AFM scan data, this is presented in the discussion section.

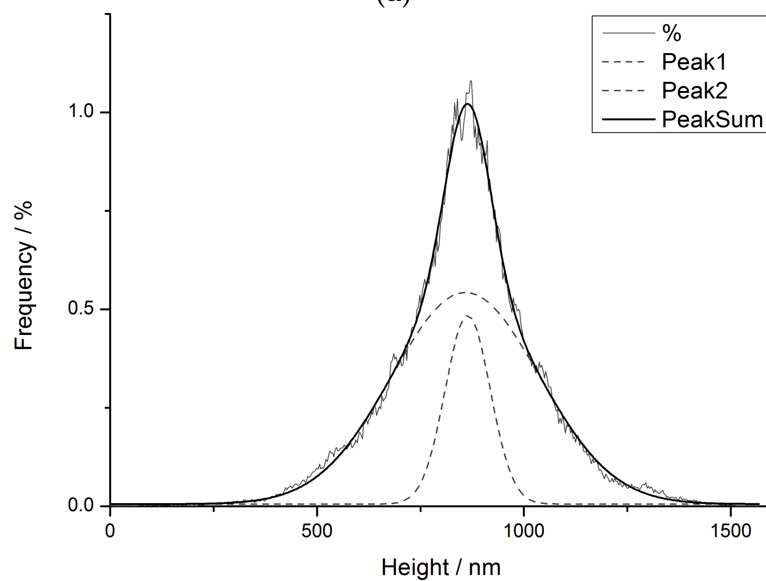
It is also possible to examine the shape of the height distribution histograms by fitting multiple Gaussian curves to the histogram peaks, examples shown in Fig. 7.10. Analysis of multi-Gaussian fits to height histograms reveal characteristic differences between grain surfaces of different origins. Care was taken to fit the histograms with the minimum number of curves required to achieve an acceptable goodness of fit. For this purpose, an acceptable R^2 value is deemed to be greater than 0.985. The shape of the height distributions depended greatly on the surface area scanned, with the most consistent results obtained from scanning surfaces away from dominant surface features, *e.g.* sharp peaks or trenches. The size of the scan had little effect on the ultimate shape of the height distributions.

7.3.3 Applicability of lateral force microscopy

One further technique that could prove useful in quantifying the surface textures of quartz sand grains is lateral force microscopy (LFM). With this form of analysis, an AFM tip is dragged laterally across the surface in contact mode and



(a)



(b)

Figure 7.10: Height distributions constructed from 20μm AFM scans of a grains from sample set 1 **(a)** and 3 **(b)**. R^2 values of 0.98935 and 0.99574 respectively.

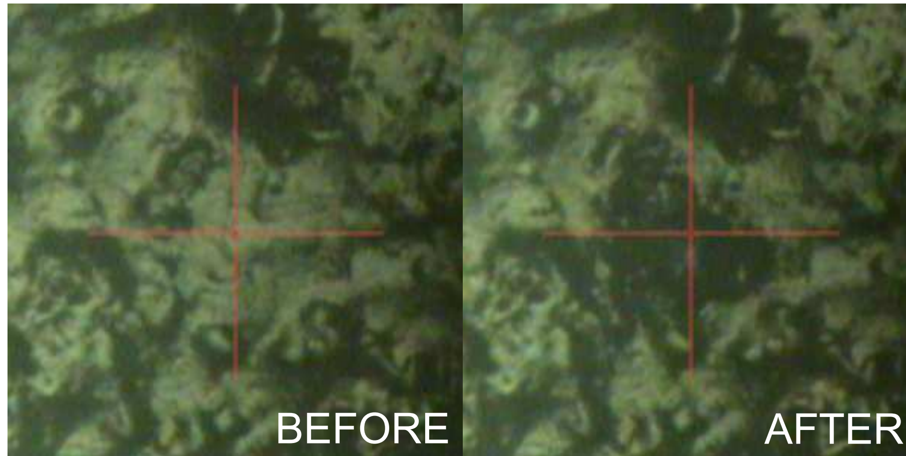


Figure 7.11: Captured image from AFM's on-board optical microscope showing damage caused by diamond tip raster scanning across a quartz grain sample surface during lateral force microscopy.

frictional forces across different surfaces are detected by measuring the twist of the cantilever. A brief overview of the technique is given in Section 2.3.1. In addition to obtaining a quantitative area roughness measurement of the surface in regular modes of operation, lateral force microscopy (in a controlled environment) has the potential to obtain frictional coefficient readings for different microscopic surface textures, thus further quantifying and enabling the categorisation of different grain types within a sample as well as from samples that are derived from different provenances.

Thus far, LFM has not proven to be a fruitful method of quartz grain analysis. Using regular Si/Si_3N_4 contact mode AFM tips for LFM resulted in the rapid erosion of the sharp tip during scanning – in some cases, the AFM tip had already been ground away by the rough textured quartz surface before a single scan could be completed. In an attempt to overcome this, diamond tips were tested. However, these harder tips resulted in an unacceptable level of sample damage. The sample surface could be visibly observed to be ground away as the tip was raster scanned in LFM mode, see Fig. 7.11.

7.3.4 Fractal nature of grain surfaces

It is well established that randomly rough surfaces can be treated as statistically self-similar fractals. Fractals usually exhibit self-similarity across scales – as the magnification increases or decreases there exists a similar, if not identical, structure. Since scanning probe microscopy techniques such as AFM or STM produce a 3D image of the surface, they are well suited to provide an indication of the fractal character of a sample. The results of the fractal analysis of rough surfaces imaged using an AFM are often used to characterise these surfaces as documented in a number of studies, including the fractal character of deposited silver [20], and chromium nitride / silicon nitride thin films [21]. There has also been work demonstrating the influence of AFM tip geometry on fractal analysis [22], and the dependence of the fractal dimension on scan conditions [23].

For some AFM scans of grains within each sample analysed for this present study, there exists a topographical self-similarity of surfaces when scanned away from any large dominant features. Figures 7.12a, 7.13a and 7.14a show 3D surface representations of a grain from sample 1 at 80μm^2, 20μm^2 and 5μm^2 respectively. The corresponding height distribution histograms, fit with multiple Gaussian curves are shown in Figs. 7.12b, 7.13b and 7.14b. This self-similar fractal nature over a given range of scales is common in fractals found in nature, making it impossible to gauge the size of the features without a scale bar.

The morphology of quartz sand grains do not appear exactly self-similar at different scales, but could easily be described as statistically self-similar (numerical or statistical measures preserved across scales), or in some grain areas, quasi-self-similar (a looser term describing approximate self-similarity).

The fractal dimension is a numerical measure that is preserved across scales for statistically self-similar fractal structures such as random fractals, *e.g.* those found in nature. To estimate the fractal dimension of a surface, various methods can be used. The box-counting method involves overlaying the fractal shape with

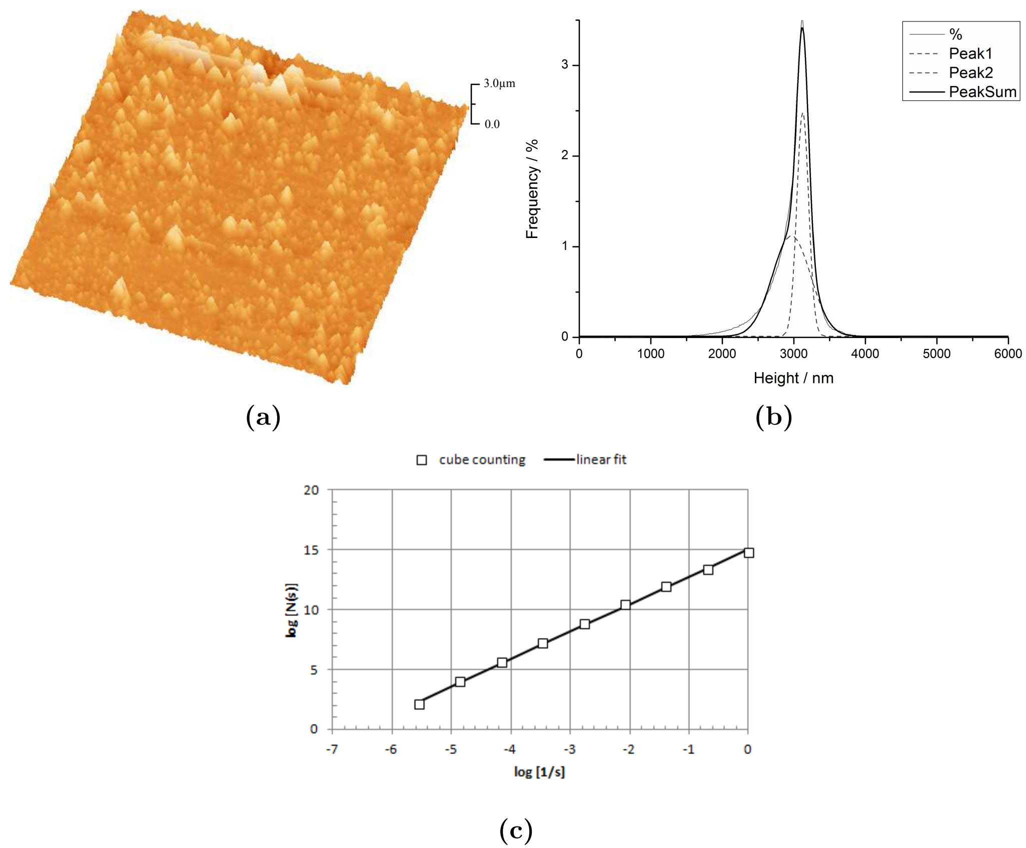


Figure 7.12: 3D surface map **(a)**, height distribution histogram **(b)**, and cube-counting estimate of fractal dimension **(c)** for 80μm^2 scan of grain surface in sample set 1 taken at 24μms^{-1} tip velocity. Fractal dimension estimated to be 2.2895.

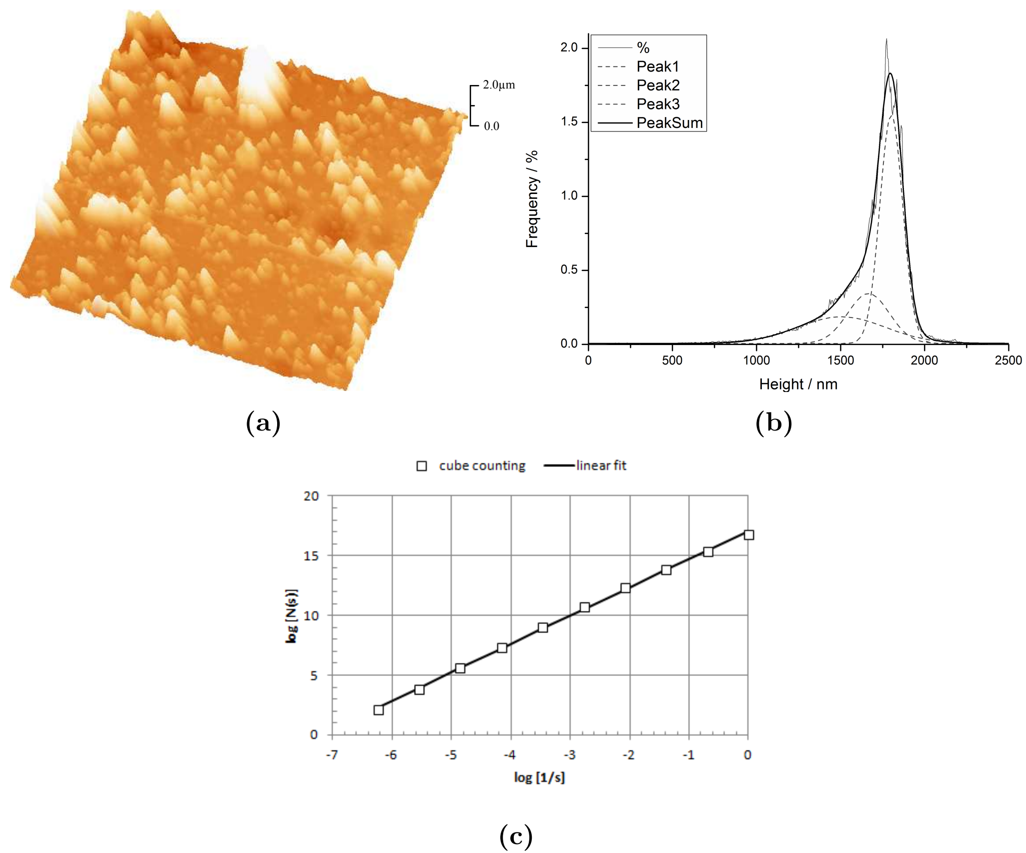


Figure 7.13: 3D surface map **(a)**, height distribution histogram **(b)**, and cube-counting estimate of fractal dimension **(c)** for 20µm^2 scan of grain surface in sample set 1 taken at 6µms^{-1} tip velocity. Fractal dimension estimated to be 2.3677.

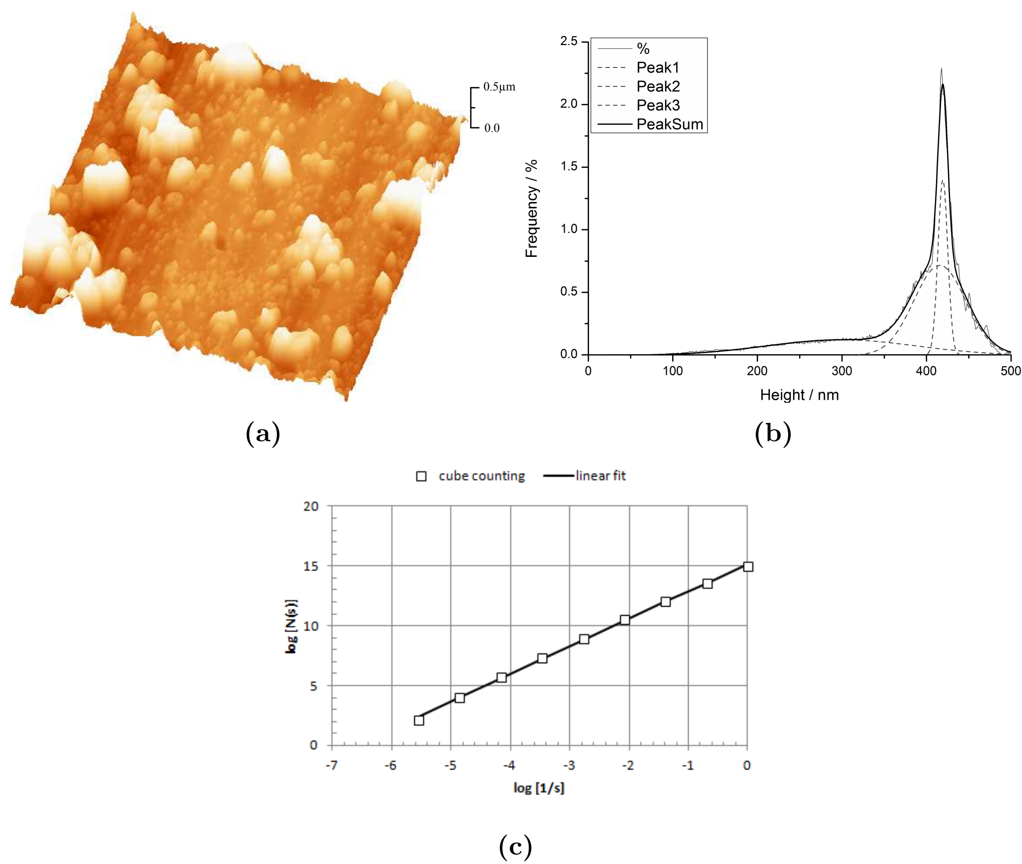


Figure 7.14: 3D surface map **(a)**, height distribution histogram **(b)**, and cube-counting estimate of fractal dimension **(c)** for $5\mu m^2$ scan of grain surface in sample set 1 taken at $2\mu ms^{-1}$ tip velocity. Fractal dimension estimated to be 2.3143.

a regular grid and counting how many boxes are required to cover the fractal's interface, *i.e.* how many boxes in the grid are intersected by the perimeter of the shape. This dimension changes as measurements are taken with reducing overlaid grid square sizes. The Minkowski-Bouligand dimension, otherwise known as the box-counting dimension, $\dim_{\text{box}}(A)$, is defined in Eq. (7.5), where: A is the 2D fractal surface; and $N(s)$ is the number of boxes of box length s.

$$\dim_{\text{box}}(A) := \lim_{s \to 0} \frac{\log N(s)}{\log(1/s)} \tag{7.5}$$

A 3D variant of this box-counting dimension known as 'cube-counting' was used to estimate the fractal dimension of the surfaces imaged by AFM. The slope of a plot of $\log(1/s)$ versus $\log N(s)$ gives us an estimate of the fractal dimension. This method allows rapid quantification of the fractal nature of surfaces.

All three scan sizes from the grain in sample set 1 have estimated fractal dimensions close to one another, between 2.3143 and 2.2895, despite the scan size reduction of 4x per step (see Figs. 7.12c, 7.13c and 7.14c).

7.4 Discussion

AFM offers many beneficial attributes as a technique to complement the forensic analysis by SEM of quartz grains. Of particular note are its abilities to image the surface topography of samples under ambient conditions, and in a non-destructive manner, but also the fact that AFM analysis requires no sample preparation, *e.g.* conductive coating to be deposited beforehand. In addition, its capability to produce quantitative results when analysing surfaces enable comparisons to be made between grains both within a sample set and between samples derived from different sites, *i.e.* from sites/scenarios corresponding to a crime scene, a suspect, articles of interest and/or alibi sites.

It is also possible with AFM analysis of the primary topographical data to construct a 3D model of the sample, allowing an examiner to more clearly vi-

sualise the surface features. It is possible to do this with an SEM, but this requires combining two or more SEM images taken at different angles and computing the 3D structure. This stereo-SEM 3D image combination works best with straight-edged, angular surfaces and suffers heavily from artefact recognition and parallax correction, making its use problematic in the analysis of quartz grains in comparison to the models derived from AFM data.

As illustrated above, along with imaging diagnostic surface features by AFM, various figures of merit can be calculated through the analysis of the topography scans which include roughness, skewness, height distribution figures, and fractal dimension which can be used to enable discrimination between quartz grain types. More research is needed to ascertain the usefulness and validity of these individual figures of merit and whether they can be interpreted individually or require cooperative figures of merit to allow for valid grain identification.

7.4.1 Discriminant analysis theory

Discriminant analysis (DA) is a statistical technique allowing the user to investigate the differences between multiple groups of objects across several variables simultaneously. The technique creates an equation (a discriminant function) based on a combination of measured variables to separate previously defined groups to the maximum extent, minimising the possibility of classifying objects into incorrect groups. The discriminant functions act to project the data onto a dimension that best discriminates between the groups.

DA is often used in a descriptive approach; to interpret group differences on the basis of the attributes of cases, assessing the adequacy of classification given the group memberships of objects. It is also common to find DA used when approaching an 'unknown classification' problem as it allows for the classification of new objects, assigning them to a number of known groups using predictive variables.

Various DA methods exist, including Fisher's linear discriminant analysis,

k-nearest neighbour analysis, and canonical discriminant analysis (also known within other fields as discriminant factor analysis or multiple discriminant analysis). Canonical discriminant analysis was deemed the most appropriate DA method for this investigation due to its focus on variable assessment.

To better characterise the use of AFM as a complimentary examination technique for quartz grain surface analysis, various topographical and statistical variables should be assessed for effectiveness in supporting the classification of grains. This focus on the variables should lead to an improved grouping ability of the data; through weighting, ineffective variables can be ignored and irrelevant variables discarded. The data should be from distinct groups, and group membership should already be known prior to DA.

There are various assumptions on DA:

- the observations are a random sample and are correctly classified in the initial description
- the number of cases per group, n_i must be at least two ($n_i >= 2$)
- the number of groups, g, must be at least two ($g >= 2$) and be defined before collecting the data
- the data is collectively exhaustive (all objects can be placed in a group)
- the groups are mutually exclusive in their classification (each object must belong to only a single group)
- the number of discriminating variables, p, must be at least one, but less than the number of total cases, n, minus two ($0 < p < (n - 2)$)
- no discriminating variable can be a linear combination of the other discriminating variables
- each group must be drawn from a population where the variables are multivariate normal (MVN)

- DA works off of the matrices used in multivariate analysis of variance (MANOVA), as such the data must not have linear dependencies (the matrices must be capable of being inverted)

The aim of DA is to determine a linear equation (akin to a regression equation) which will predict which group a case belongs to. This takes the form of the canonical discriminant function shown in Eq. (7.6).

$$f_{km} = u_0 + u_1 X_{1km} + u_2 X_{2km} + ... + u_p X_{pkm} \tag{7.6}$$

Where: f_{km} is the value (score) on the canonical discriminant function for case m in group k; X_{ikm} is the value on discriminating variable X_i for case m in group k; and u_i is the coefficient which produces the desired characteristic of the function.

A set of discriminant functions are derived to provide maximum multiple correlation with the groups. The first function is built to maximise group differences. The second function is then built to be orthogonal to the first function while still maximising group differences, *i.e.* they provide independent, uncorrelated, non-overlapping contributions to discrimination between groups. The process of deriving uncorrelated functions is repeated until the number of functions derived is equal to either the total number of groups minus one or the total number of discriminating variables, whichever is smaller ($\min(g-1, p)$). Successive functions are less effective than the last, and thus only the first few are typically required for analysis.

To obtain the canonical discriminant function, a set of squares and cross products (SSCP) matrices are constructed:

- Total SSCP matrix ($\mathbf{T}$)
- Within-group SSCP matrix ($\mathbf{W}$)
- Between-group SSCP matrix ($\mathbf{B}$)

Each element of the total SSCP Matrix:

$$t_{ij} = \sum_{k=1}^{g} \sum_{m=1}^{n_k} (X_{ikm} - X_{i..})(X_{jkm} - X_{j..}) \tag{7.7}$$

Each element of the within-group SSCP Matrix:

$$w_{ij} = \sum_{k=1}^{g} \sum_{m=1}^{n_k} (X_{ikm} - X_{ik.})(X_{jkm} - X_{jk.}) \tag{7.8}$$

Where: g is the number of groups; n_k is the number of cases in group k; $n.$ is the number of cases over all groups; X_{ikm} is the value of variable i for case m in group k; $X_{ik.}$ is the mean value of variable i for cases in group k; and $X_{i..}$ is the mean value of variable i for all cases.

Once the total and within-group SSCP matrices are obtained, the between-group SSCP matrix can be calculated as: **B**=**T**-**W**. When there are no differences between the group centroids (mean vectors of each group), **W**=**T**. The extent that they differ defines the distinctions among observed variables.

Once the between-group and within-group matrices are obtained, the eigenvalues and eigenvectors can be taken and the solutions (v_i) to the following equations found:

$$\begin{aligned}
\sum b_{1i} v_i &= \lambda \sum w_{1i} v_i \\
\sum b_{2i} v_i &= \lambda \sum w_{2i} v_i \\
&\vdots \\
\sum b_{pi} v_i &= \lambda \sum w_{pi} v_i
\end{aligned} \tag{7.9}$$

Once the λ and v_i parameters are found, they are converted into the weights for the discriminant functions.

$$u_i = v_i\sqrt{n. - g} \tag{7.10}$$

$$u_0 = -\sum_{i=1}^{p} u_i X_{i..} \tag{7.11}$$

Groups are scored using these weights and can be positive or negative. A positive object is said to be "high on a dimension"; inversely, a negative object is "low on a dimension". The units are in standard deviation units of the discriminant space. The group means of each of the discriminant functions is most often used in classifying observations.

There are several methods available to classify objects; all are based on the distance of the object from group centroids. The easiest method is to classify all objects into groups according to their nearest group centroids, and then to assess how well the original groups can be recovered.

7.4.2 Discriminant analysis output explanation

The results output from DA are typically given as four individual tables, often accompanied by a canonical discriminant function plot, but many others are available. This section will explain the results output in these tables for reference later in the relevant results sections.

Eigenvalues

The eigenvalues column gives the eigenvalues calculated from the product of the inverse within-group and between-group SSCP matrices associated with each of the canonical linear discriminant functions (equations) produced. The magnitude of each function's eigenvalue gives an indication of the extent of that function's discriminating ability, and is related to the canonical correlations.

The percentage of variance for each function gives the proportion of discriminating ability that is attributed to each function. This is calculated as the

proportion of the function's eigenvalue to the sum of all the functions' eigenvalues. The next column gives the cumulative percentage of discriminating ability.

The final column gives the canonical correlation, the multiple correlation between the predictor variables and the discriminant function. If there is only a single function being analysed, it provides a value indicative of the fit of the model. The percentage of variability explained by the model can be obtained by squaring the canonical correlation value.

Wilks' lambda

Wilks' Lambda is a multivariate statistic giving the proportion of total variability not explained by the discriminant function. Since the percentage of variability explained by the model is given by the canonical correlation squared (CC^2), Wilks' Lambda is simply calculated as the product of the values of $1 - CC^2$. It should be mentioned that a test of '1 through 2' is Wilks' Lambda testing of the first two canonical correlations used, whereas a test of '2' would just be testing the second function's canonical correlation. It also indicates the significance of the discriminant function.

The null hypothesis is the default position - that there is no relationship between measured phenomena. Thus, to 'reject the null hypothesis' is to accept a statistical relationship between them. In this case, the null hypothesis is that the function, and all subsequent functions that follow, have no discriminating ability. The Chi-square statistic is used to test this null hypothesis. The 'df' column gives the effective degrees of freedom for each function, a value based on the number of discriminant variables and groups. The Chi-squared statistic may be compared to a Chi-squared distribution using the degrees of freedom stated.

Significance is an indicator of whether the observations reflect a pattern or are just down to random chance. A result is statistically significant if it were expected to arise by chance very rarely. The significance column gives the alpha, which the p-value associated with the Chi-square statistic is evaluated against.

If the p-value is smaller than alpha, the null hypothesis is rejected and the results are deemed statistically significant. Commonly used significance levels are 95% (0.05) and 99% (0.01), but this can vary depending on the situation, a p-value less than 0.001 translates as a 0.01% or 1-in-1000 chance of the result being down to chance, and $p < 0.000$ indicates a highly significant result, and translates as having a 'less than 1-in-1000 chance' of the result being random.

Functions at group centroids

These are the means (group centroids) of the discriminant function scores by group for each function calculated. The function scores have a mean of zero, and this can be verified by taking the sum of the multiple of the number of cases and the group function mean scores.

Classification results

Reading horizontally across each row, the predicted group memberships for each original group can be seen. These frequencies indicate how many of the original group cases fall into each of the original groups after classification according to the discriminant analysis functions. These predicted group membership frequencies are given in percentages below the frequency data.

Finally, in table footnote [c] the classification results table contains a figure for the percentage of original grouped cases correctly identified by the canonical discriminant analysis, *i.e.* how well the original groups could be recovered.

7.4.3 Canonical discriminant functions from original sample sets

In the initial study, when canonical discriminate function analysis was undertaken of these measures for the quartz grains analysed, it was possible to distinguish between the quartz grains derived from the three samples at the 99% significance level on the basis of the height distribution statistical measures, (Fig. 7.15 and complementary Table 7.2 (Wilks' Lambda = 0.614, $p = 0.000 <$

Table 7.2: Statistical output from canonical discriminant function analysis of AFM topography and derived height distribution statistical (skewness, kurtosis) data.

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	0.479[a]	82.4	82.4	0.569
2	0.102[a]	17.6	100	0.304

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 2	0.614	47.635	8	0.000
2	0.907	9.476	3	0.024

Functions at group centroids

Sample Group	Function[b]	
	1	2
sample 1	0.902	-0.157
sample 2	-0.181	0.409
sample 3	-0.773	-0.315

Classification Results[c]

	Sample Group	Predicted Group Membership			Total
		sample 1	sample 2	sample 3	
Count	sample 1	27	1	6	34
	sample 2	10	8	19	37
	sample 3	3	2	26	31
%	sample 1	79.4	2.9	17.6	100
	sample 2	27	21.6	51.4	100
	sample 3	9.7	6.5	83.9	100

a. First 2 canonical discriminant functions were used in the analysis.
b. Unstandardised canonical discriminant functions evaluated at group means.
c. 59.8% of original grouped cases correctly classified.

0.01)); the fractal measures, (Fig. 7.16 and Table 7.3 (Wilks' Lambda = 0.735, $p = 0.000 < 0.01$)); as well as on the basis of the statistical roughness measures, (Fig. 7.17 and Table 7.4 (Wilks' Lambda = 0.604, $p = 0.000 < 0.01$)).

7.4.4 Final canonical discriminant function

With additional data taken from a new fourth sample set obtained from a Skegness beach murder investigation, along with new scans of grains from the original sample sets, the canonical discriminant analysis was recomputed over all statistical measures. The results are shown in Table 7.5, along with a canonical discriminant function plot of these final functions, shown in Fig. 7.18. The inclusion of all variables, new scans, and a fourth sample set resulted in a far better fitting canonical discriminant function with a much lower Wilks' Lambda

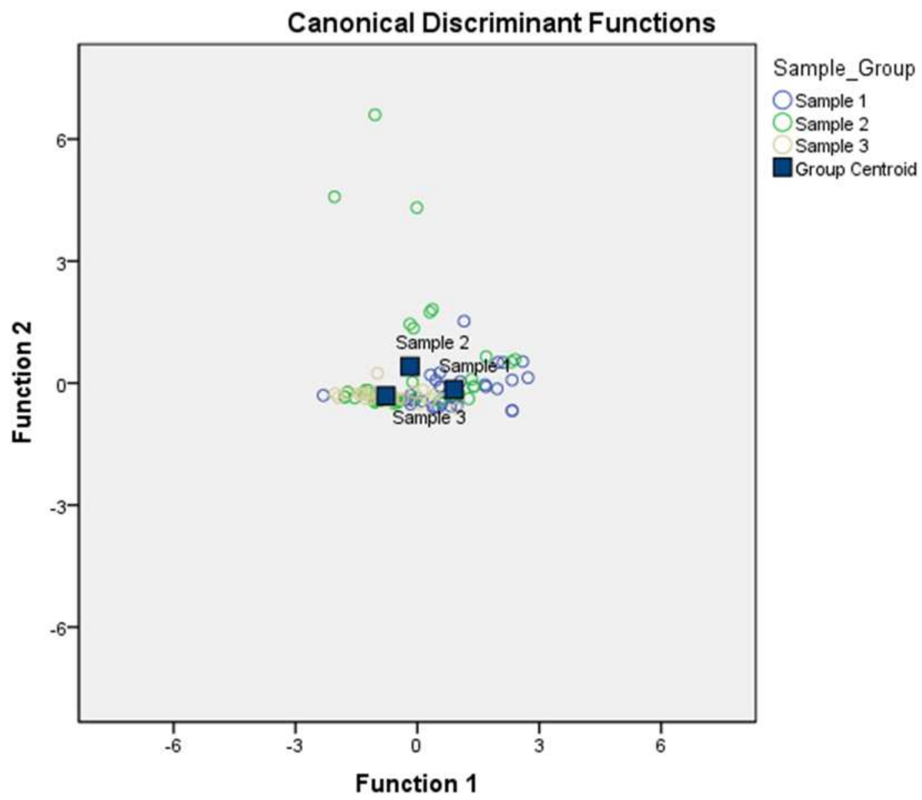


Figure 7.15: Canonical discriminant function plot of AFM topography and derived height distribution statistical measures (skewness, kurtosis) from AFM scans of the three sample sets.

Table 7.3: Statistical output from canonical discriminant function analysis of AFM fractal dimension data.

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	0.278[a]	81.2	81.2	0.467
2	0.065[a]	18.8	100	0.246

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 2	0.735	33.74	8	0.000
2	0.939	6.858	3	0.077

Functions at group centroids

Sample Group	Function[b]	
	1	2
sample 1	0.703	0.079
sample 2	-0.495	0.226
sample 3	-0.201	-0.381

Classification Results[c]

	Sample Group	Predicted Group Membership			Total
		sample 1	sample 2	sample 3	
Count	sample 1	25	6	8	39
	sample 2	13	15	14	42
	sample 3	7	8	18	33
%	sample 1	64.1	15.4	20.5	100
	sample 2	31	35.7	33.3	100
	sample 3	21.2	24.2	54.5	100

a. First 2 canonical discriminant functions were used in the analysis.
b. Unstandardised canonical discriminant functions evaluated at group means.
c. 50.9% of original grouped cases correctly classified.

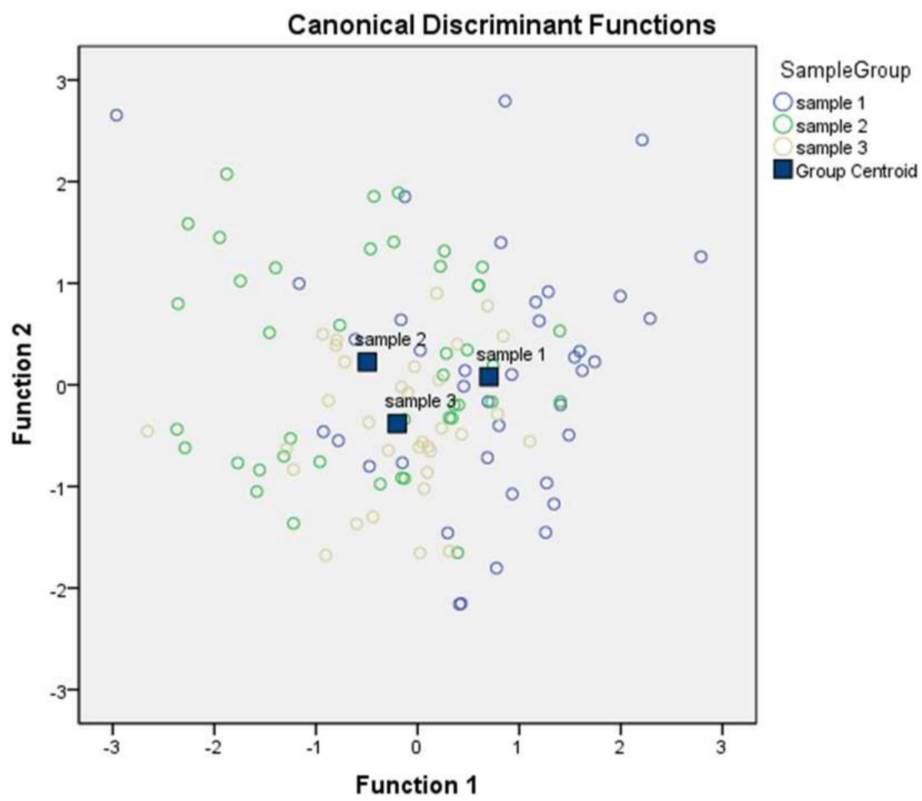


Figure 7.16: Canonical discriminant function plot of fractal dimension figures of merit from AFM scans of the three sample sets.

Table 7.4: Statistical output from canonical discriminant function analysis of AFM statistical 'roughness' (S_a, S_q, S_{sk}, S_{ku}) data.

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	0.406[a]	69.5	69.5	0.537
2	0.178[a]	30.5	100	0.389

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 2	0.604	55.254	8	0.000
2	0.849	17.942	3	0.000

Functions at group centroids

Sample Group	Function[b]	
	1	2
sample 1	-0.422	0.505
sample 2	-0.382	-0.483
sample 3	0.985	0.018

Classification Results[c]

	Sample Group	Predicted Group Membership			Total
		sample 1	sample 2	sample 3	
Count	sample 1	22	10	7	39
	sample 2	9	25	8	42
	sample 3	6	4	23	33
%	sample 1	56.4	25.6	17.9	100
	sample 2	21.4	59.5	19	100
	sample 3	18.2	12.1	69.7	100

a. First 2 canonical discriminant functions were used in the analysis.
b. Unstandardised canonical discriminant functions evaluated at group means.
c. 61.4% of original grouped cases correctly classified.

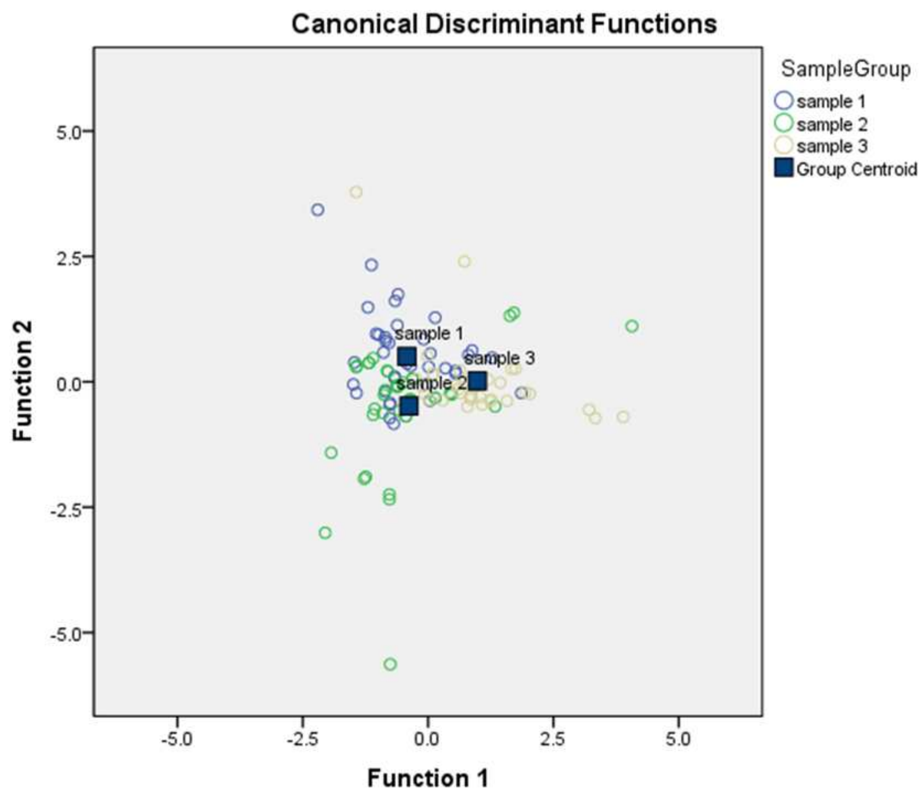


Figure 7.17: Canonical discriminant function plot of statistical 'roughness' measures (S_a, S_q, S_{sk}, S_{ku}) of AFM scans of the three sample sets.

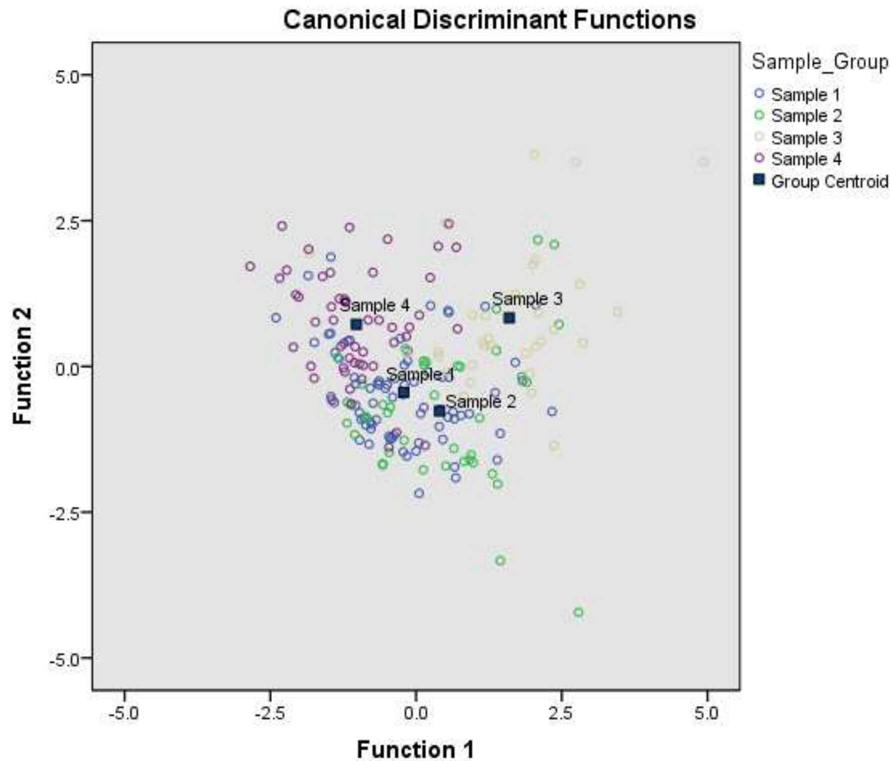


Figure 7.18: Canonical discriminant function plot of all derived statistical figures of merit for 4 sample sets.

value of 0.336 for a significance $p = 0.000 < 0.01$. This degree of discrimination shows a great deal of promise for this technique as a quantitative supplement to traditional SEM grain examination.

Clearly, the samples here are derived from highly distinct provenances, and for the forensic capability of these figures of merit to be established further experimental work needs to be undertaken using samples that can be considered to have ecological validity for forensic casework [24]. As grains belonging to more sample sets of known provenance are examined using an AFM, the addition of their data will only add to the accuracy of the multivariate discriminant function. Only a handful of statistical measures were taken during this investigation, as new measures are found and tested for their discriminatory abilities, they too can be added to the discriminant function.

Table 7.5: Statistical output from canonical discriminant function analysis of all derived statistical figures of merit for 4 sample sets.

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	0.752[a]	54.7	54.7	0.655
2	0.449[a]	32.7	87.4	0.556
3	0.173[a]	12.6	100	0.384

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 3	0.336	211.680	30	0.000
2 through 3	0.588	102.890	18	0.000
3	0.852	30.998	8	0.000

Functions at group centroids

Sample Group	Function[b]		
	1	2	3
sample 1	-0.212	-0.442	0.442
sample 2	0.404	-0.765	-0.635
sample 3	1.599	0.833	0.114
sample 4	-1.023	0.721	-0.218

Classification Results[c]

	Sample Group	Predicted Group Membership				Total
		sample 1	sample 2	sample 3	sample 4	
Count	sample 1	43	13	7	13	76
	sample 2	12	24	4	1	41
	sample 3	5	1	26	1	33
	sample 4	7	1	5	39	52
%	sample 1	56.6	17.1	9.2	17.1	100
	sample 2	29.3	58.5	9.8	2.4	100
	sample 3	15.2	3.0	78.8	3.0	100
	sample 4	13.5	1.9	9.6	75.0	100

a. First 3 canonical discriminant functions were used in the analysis.
b. Unstandardised canonical discriminant functions evaluated at group means.
c. 65.3% of original grouped cases correctly classified.

7.5 Conclusions

This chapter presents the findings of an initial study to assess the capabilities and potential for the use of AFM as an independent verification of quartz grain type classifications for forensic discrimination purposes. The quantifiable measures that the AFM is able to provide, in addition to the additional imaging capabilities, provide independent measures of the quartz surface textures that can be used in combination with the morphological classification of quartz grain types. Given the preliminary nature of this study, only four samples from clearly distinct provenances were investigated and further, more extensive investigations incorporating grains of many more origins and the in-depth examination of their 'diagnostic' surface features must be performed to fully appreciate the full potential of this technique. From the analysis undertaken for this study it appears that the quantifiable figures of merit that can be derived for quartz surfaces can be utilised to good effect in the discrimination of different quartz grain types. Further work should focus on samples from a single origin that have suites of multiple grain types present, and also on the capacity of the AFM to distinguish between grains that have similar morphological features that have derived from different provenances (building on the work of Newell *et al.* [8]).

In conclusion, AFM analysis can provide topographical data from the grain surface that enables statistical analysis, 3D reconstruction and quantitative assessments of the microscopic surface textures. Such analysis clearly enhances and builds upon the morphological assessments of the surface textures that can be achieved by SEM, and demonstrates the potential for independent corroboration of quartz types. Such independent corroboration provides a means of not only strengthening the interpretations reached from quartz grain surface texture analysis, but enabling this form of trace evidence to provide more robust and verifiable intelligence, and potentially evidence, in a forensic context (in a similar manner to initial ion beam analysis studies [9]). As with many forms of analysis

in this domain, AFM is not a 'one-stop' technique for the examination of quartz grains, but appears to offer a range of supplementary quantitative data about the samples that provide additional discriminatory information for discerning between samples derived from different provenances.

The quantifiable measure of quartz grain surface textures that is achievable with AFM opens up a number of possibilities for forensic quartz grain surface texture analysis in terms of providing a corroborative independent verification of quartz type classifications. It is, however, important to consider that although the optical microscope on the AFM allows for accurate placement of the tip for scanning a specific area of the grain, this is still far reduced from the imaging size capability of an SEM. The limited x-y scan area of the AFM restricts the maximum scan size to approximately $100\mu m^2$ and the maximum z-distance to $10\mu m$, and this should be taken into account in the development of AFM and in assessing the best methods of employing its capabilities as a corroborative technique in combination with SEM analysis within forensic contexts for quartz grain analysis.

This study highlights the potential for AFM analysis in the forensic discrimination of quartz grains. It offers numerous statistical methods to discriminate between grain surface textures, which combined with the possibility for a more traditional qualitative examination, allows an approach previously unavailable with traditional examination, that of an automated database to compile and generate reports. A likely grain analysis system could take the form of a procedure whereby after obtaining an 'acceptable' scan, the data are analysed automatically and compared with figures from similar grains and compiled into an automated report that could provide a means for comparing quartz grains derived from samples of different provenances for forensic reconstructions. This report could contain various items:

- 2D topography scan taken of grain

- Qualitative comparison of diagnostic surface features to ascertain history of grain
- Quantification of surface roughness and other simple statistical figures to categorise grain
- Statistical analysis of height distributions *e.g.* featuring multi-Gaussian curve fitting to further discriminate between grains
- Computing the fractal dimension of the surface for additional discrimination

There is clear potential for such an approach to be tailored to a particular forensic case where soil/sediment samples collected from different locations and items of interest are available. The benefits would lie not only in providing a screening of a large number of grains to enable subsequent in depth analysis of the grains from samples that could not be excluded in the first screen, but this approach would reduce operator time, and therefore the cost of such analyses. Increasing the analysis speed and reducing operator time would enable far more samples to be analysed and make the approach more widely available to a broader range of forensic investigations. These results, whilst only providing an initial capability study, nevertheless signal the potential for AFM to provide a highly valuable and usable additional tool for soil/sediment analyses in forensic enquiries.

Bibliography

[1] D. Konopinski, S. Hudziak, R. Morgan, P. Bull, and A. Kenyon, "Investigation of quartz grain surface textures by atomic force microscopy for forensic analysis," *Forensic Science International*, vol. 223, no. 1, pp. 245–255, November 2012.

[2] A. Ruffell and J. McKinley, *Geoforensics.* John Wiley & Sons, 2008.

[3] K. Ritz, L. Dawson, and D. Miller, Eds., *Criminal and Environmental Soil Forensics.* Springer, 2009.

[4] P. A. Bull and R. M. Morgan, "Sediment fingerprints: a forensic technique using quartz sand grains," *Science & Justice*, vol. 46, no. 2, pp. 107–124, Apr. 2006. [Online]. Available: http://www.scienceandjusticejournal.com/article/S1355-0306(06)71581-7

[5] R. M. Morgan, P. Wiltshire, A. Parker, and P. A. Bull, "The role of forensic geoscience in wildlife crime detection," *Forensic Science International*, vol. 162, no. 1, pp. 152–62, Oct. 2006.

[6] R. M. Morgan, M. Little, A. Gibson, L. Hicks, S. Dunkerley, and P. A. Bull, "The preservation of quartz grain surface textures following vehicle fire and their use in forensic enquiry," *Science & Justice*, vol. 48, no. 3, pp. 133–140, Sep. 2008.

[7] R. M. Morgan and P. A. Bull, "Forensic geoscience and crime detection Identification , interpretation and presentation in forensic geoscience," *Minerva Medicolegal*, vol. 127, no. 2, pp. 73–89, Jun. 2007. [Online]. Available: http://www.minervamedica.it/it/riviste/minerva-medicolegale/articolo.php?cod=R11Y2007N02A0073

[8] A. J. Newell, R. M. Morgan, L. D. Griffin, P. a. Bull, J. R. Marshall, and G. Graham, "Automated Texture Recognition of Quartz Sand Grains for

Forensic Applications," *Journal of Forensic Sciences*, vol. 57, no. 5, pp. 1285–1289, Sep. 2012.

[9] M. J. Bailey, R. M. Morgan, P. Comini, S. Calusi, and P. A. Bull, "Evaluation of particle-induced X-ray emission and particle-induced γ-ray emission of quartz grains for forensic trace sediment analysis," *Analytical Chemistry*, vol. 84, no. 5, pp. 2260–2267, Mar. 2012.

[10] J. R. Smith, "A quantitative method for analysing AFM images of the outer surfaces of human hair," *Journal of Microscopy*, vol. 191, no. 3, pp. 223–228, Sep. 1998.

[11] S. P. Gurden, V. F. Monteiro, E. Longo, and M. M. Ferreira, "Quantitative analysis and classification of AFM images of human hair," *Journal of Microscopy*, vol. 215, no. 1, pp. 13–23, Jul. 2004.

[12] E. Canetta, K. Montiel, and A. K. Adya, "Morphological changes in textile fibres exposed to environmental stresses: atomic force microscopic examination," *Forensic Science International*, vol. 191, no. 1-3, pp. 6–14, Oct. 2009.

[13] S. Kasas, A. Khanmy-Vital, and G. Dietler, "Examination of line crossings by atomic force microscopy," *Forensic Science International*, vol. 119, no. 3, pp. 290–298, Jul. 2001.

[14] R. H. Bremmer, K. G. de Bruin, M. J. C. van Gemert, T. G. van Leeuwen, and M. C. G. Aalders, "Forensic quest for age determination of bloodstains," *Forensic Science International*, vol. 216, no. 1-3, pp. 1–11, Mar. 2012.

[15] S. Strasser, A. Zink, G. Kada, P. Hinterdorfer, O. Peschel, W. M. Heckl, A. G. Nerlich, and S. Thalhammer, "Age determination of blood spots in forensic medicine by force spectroscopy," *Forensic Science International*, vol. 170, no. 1, pp. 8–14, Jul. 2007.

[16] E. Canetta and A. K. Adya, "Atomic force microscopic investigation of commercial pressure sensitive adhesives for forensic analysis," *Forensic Science International*, vol. 210, no. 1-3, pp. 16–25, Jul. 2011.

[17] C. Bersellini, L. Garofano, M. Giannetto, F. Lusardi, and G. Mori, "Development of latent fingerprints on metallic surfaces using electropolymerization processes," *Journal of Forensic Sciences*, vol. 46, no. 4, pp. 871–877, Jul. 2001.

[18] National Research Council of the National Academies, *Strengthening Forensic Science in the United States: A Path Forward*. The National Academies Press, 2009. [Online]. Available: http://www.nap.edu/catalog/12589.html

[19] The Law Commission, *Consultation Paper No 190: The admissibility of expert evidence in criminal proceedings in England and Wales - A new approach to the determination of evidentiary reliability*. The Law Commission, 2010, no. 190. [Online]. Available: http://www.justice.gov.uk/lawcommission/docs/cp190_Expert_Evidence_Consultation.pdf

[20] C. Douketis, Z. Wang, T. L. Haslett, and M. Moskovits, "Fractal character of cold-deposited silver films determined by low-temperature scanning tunneling microscopy," *Physical Review B*, vol. 51, no. 16, pp. 11 022–11 031, 1995.

[21] W. Zahn and A. Zösch, "Characterization of thin-film surfaces by fractal geometry," *Fresenius' Journal of Analytical Chemistry Journal of Analytical Chemistry*, vol. 358, no. 1, pp. 119–121, 1997.

[22] A. Mannelquist, N. Almqvist, and S. Fredriksson, "Influence of tip geometry on fractal analysis of atomic force microscopy images," *Applied Physics A: Materials Science & Processing*, vol. 66, no. 0, pp. S891–S895, 1998.

[23] W. Zahn and A. Zösch, "The dependence of fractal dimension on measuring conditions of scanning probe microscopy," *Fresenius' Journal of Analytical Chemistry*, vol. 365, no. 1, pp. 168–172, 1999.

[24] R. M. Morgan, J. Cohen, I. McGookin, J. Murly-Gotto, R. O'Connor, S. Muress, J. Freudiger-Bonzon, and P. a. Bull, "The relevance of the evolution of experimental studies for the interpretation and evaluation of some trace physical evidence," *Science & Justice*, vol. 49, no. 4, pp. 277–285, Dec. 2009.

Chapter 8

Conclusions and future work

This chapter presents a summary of the significant results and conclusions of the thesis, along with application-specific contributions to the fields of microscopy and forensics. Alongside this will be a discussion of the main obstacles that have yet to be overcome and possible solutions to these issues. Finally, suggestions are made for further research and development areas to fully realise these two forensic methods, and extend this work into the future.

8.1 Conclusions

The overall aim of this thesis has been to establish the potential of two previously undeveloped forensic science techniques utilising the AFM. The first involved the development and characterisation of a forensically sound method for the sample processing and extraction of data from potentially damaged SIM card evidence. The second project undertaken was to assess the complimentary potential of the AFM as a quantitative investigative technique for the examination of quartz grain surface textures for forensic analysis.

8.1.1 SIM card processing and data extraction

For the SIM card project, a process has been successfully developed to accurately expose the underside of the embedded EEPROM/flash memory arrays within smart card microcontrollers. This process was developed following various previously published studies, including an earlier study by De Nardi *et al.* [1–3]. Ultimately, this methodology incorporates many techniques drawn from other fields, successfully applying them towards the development of a forensically sound sample preparation methodology. Techniques from such interdisciplinary fields as failure analysis, materials science, and semiconductor fabrication have all been fundamental to the final method.

The chip extraction and decapsulation process was successfully developed for use on potentially damaged SIM cards and other smart card devices. Testing various conclusions and suggested decapsulation methods from numerous published sources in the field of failure analysis, most often dealing with dual in-line (DIL) packaged microcontrollers, led to the successful development of a two-part decapsulation process. This process involves the application of heat to loosen the hot melt adhesive holding the SIM card chip module to the main card body. This is followed by submersing the chip module in heated fuming nitric acid to dissolve the epoxy glob top surrounding the microcontroller.

Following the successful application of this fuming nitric acid chip extraction process, several pre-programmed SIM cards were decapsulated. These samples were baked to various temperatures between 100 and 600°C in a 'stabilisation bake' investigation with two goals: to assess the adequacy of the Kelvin probe apparatus as an initial sample analysis tool; and to better establish the data retention characteristics of SIM cards when exposed to elevated temperatures. Kelvin probe measurements were taken before and after baking each chip, but the results showed no significant correlation between the surface potential and the temperature it has been exposed to, *i.e.* the level of charge remaining within

the floating gates. The second investigation was a follow-up to a brief, but original study by Jones & Kenyon [4, 5], and produced a particularly significant result. Uncorrupted data could be successfully retrieved after rewiring from some devices heated to 500°C for 20 minutes in a furnace, this temperature exceeds all but the short-lived living room desk-height peak temperature of 540°C encountered in Purporti's full-scale house fire investigation [6].

Following successful re-encapsulation of the extracted dies within chemically-resistant epoxy resin casts for protection, the bulk silicon is thinned to <40μm thickness by mechanical lapping with fixed abrasive particles. This machining process leaves the silicon surface scratched and scored from the crude abrasive material removal. To remove these surface aberrations, the sample is lapped with progressively finer abrasive papers, eventually moving on to free-flowing diamond abrasive pastes. To more accurately characterise these processes and the effects upon the subsequent etching uniformity, various samples were lapped and polished, and etched with TMAH. The results of this investigation determined that polishing to a minimum of 1μm diamond grit size is necessary to achieve R_{a} <10nm, but polishing with abrasives below 1μm makes little difference to etching uniformity.

With the bulk silicon reduced to an appropriate thickness, the final stage was to chemically etch away the remaining silicon using TMAH. This organic-ballasted ammonium hydroxide variant is an anisotropic silicon etchant used in semiconductor and MEMS fabrication processes. To define the conditions required for a forensically sound process, various TMAH etching experiments were conducted. These experiments assessed typical etching conditions, and the resultant etch rates and surface roughnesses – these roughnesses being a statistical descriptor of the overall etch uniformity. The resulting developed process was to etch the sample in a solution of TMAH 25wt% Aq. with IPA 10vol% heated in a water bath to a temperature not exceeding 80°C, or if using a hotplate with a magnetic stirrer to set the maximum temperature no higher than

75°C. This process should take between 2 and 2.5 hours to complete, and results in the safe removal of the silicon substrate. The overall accomplishment of these sample preparation investigations has been the full development of a suitable methodology for the extraction and processing of SIM card microcontrollers, safely exposing the <10nm tunnel oxide layer while avoiding damaging the tunnel oxide layer which would destroy the data.

Preparing samples in accordance with this successfully developed process selectively removes the bulk silicon while leaving any SiO_2 structures relatively untouched. The underside of this circuit is now prepared, with the thin tunnel oxide regions adjacent to the floating gates exposed. The quantity of injected electrons within these floating gates determines the threshold voltages of the transistors, and hence the cells' logical states. With the tunnel oxides exposed, the use of multiple-pass electrical SPM modes has been explored as a means of probing the relative level of stored charge held within the electrically isolated floating gates.

While the application of electrical SPM has not (yet) been fully realised for use in examining SIM card memory arrays, it has certainly produced some significant results. The first SPM technique tested was EFM. This technique was stated to be unsuitable for such an application in previous studies by De Nardi *et al.* [1–3]. Operating similar to SKPM but without the tip bias feedback system, this mode has been shown to be less suitable for this application, in particular being far more susceptible to interference from nearby topography, in this case the adjacent field oxide structures.

Unfortunately, with far fewer injected electrons in modern floating gate transistors, coupled with a reduced fabrication process leading to more substantial topographical artefacts, SKPM scans were unable to vividly distinguish between charged and uncharged gates in a similar manner to what De Nardi *et al.* had previously published. Their samples consisted of much larger featured, lower capacity, dedicated EEPROM memory chips fabricated with a process node first

introduced 18 years ago. Surface potential scans of their samples were quite striking, programmed gates containing injected charge producing distinct bright regions clearly showing a distinction between these and the darker erased cells nearby.

Despite not exhibiting the distinct bright regions similar to De Nardi's results, SKPM was successfully performed on pre-programmed processed SIM card microcontrollers. The results have shown this technique to be capable of quantitatively probing the floating gates beneath the tunnel oxide layer, or more accurately: the capacitively-coupled charges mirrored on the tunnel oxide-air interface. The data has been analysed and plotted in a histogram form, clearly showing the difference between charge levels of gates pertaining to logical 0 and 1 states. These Gaussian-fitted distributions are reminiscent of the stored charge voltage windows commonly encountered with EEPROM/flash memory cell literature.

Unfortunately, the surface charges on the tunnel oxide-air interface imaged by SKPM are of a lower proportion to the actual level of charge within the floating gate. Combining this with fewer electrons injected to swap logical states in modern devices, and the effect on the charged gate (logical 0) surface potential distribution is clear – it is shifted towards the uncharged (logical 1) distribution. For the results shown in this study, this produces a substantial overlap between the two distributions, resulting in a high degree of uncertainty in the SKPM-measured logical states of the cells.

8.1.2 Quartz grain surface texture analysis

The results in this thesis show the AFM has been successfully applied to the novel forensic examination of quartz sand grain surface textures. This work was certainly one of the first few times AFM has ever been performed on quartz sand grains [7], but almost certainly the first time (to the author's knowledge) that it has been used to examine sand grains in a forensic context. Aimed

at supplementing traditional SEM analysis of grain micro-textures, the AFM proved capable of imaging grain surface topography and producing three dimensional reconstructions of the grain surface. Topographical statistical measures and height distributions were calculated and compared, followed by a discussion on the applicability and hazards of lateral force microscopy and other contact mode AFM/SPM techniques. This was followed by a look at the statistical self-similarity of the fractal nature of micrometre-scale grain surfaces. Multiple grains from four sample sets were ultimately examined, and readily available statistical measures were computed and gathered from the grain surface topographies.

Forensic science is a broad term used to describe the application of numerous analytical techniques used to generate results of evidentiary value to judicial investigations. These analyses are often used to compare samples from different sources in an exclusionary manner. The basis for such analysis lies in the examination of previously proven sample characteristics, and it is this critical and quantitative assessment of said characteristics which form the foundation of any forensic science technique – be it novel or established. To this end, various statistical measures were used to develop a multivariate function using canonical discriminant analysis. This allows sand grains to be classified based on grain provenance, determined statistically on AFM-measured quantifiable figures of merit. The final results from examining numerous grains from 4 sample sets were that 65.3% of original grouped cases were able to be correctly classified, with the first 3 discriminant functions used in the analysis. The functions' Wilks' Lambda value was 0.336 with a significance $p = 0.000 < 0.01$. This excellent discriminatory ability shows the AFM has great promise as a supplementary quantitative technique to existing qualitative forensic SEM analysis.

8.2 Future work and further applications

The contributions made in this thesis provide an excellent foundation for the future development of both the recovery of data from damaged SIM cards and the examination of quartz sand grains. This section will focus on some possible future avenues for research and development within these two areas, as well as an overview of potential further applications should either technique become fully established. As is often the case with research, more questions than answers are produced, and the forensic sciences are no exception – for every developed method there is a high level of scrutiny over where and when it can be applied.

At this stage there are several avenues available for further development, but the most obvious is the complete development and rigorous characterisation of SKPM. The work outlined in this thesis has rightly shown the potential of these microscopy techniques, but also their limitations with modern devices. There is still the possibility that these limitations may be overcome, allowing this technique to better discern between individual logical states. This would produce results of a greater significance than the current overlapping distributions of surface potentials. If this SPM analysis were to be perfected, researchers would be faced with further obstacles to overcome before the technique were fully realised as a data extraction tool for forensic examiners.

The first task would be to piece together multiple dual-channel topography/potential scans. These data files consist of a continuous stream of coordinates gathered in raster fashion, with corresponding measurements for each data channel. Overlapping scans can be joined using a variety of methods, and software packages are already available to construct large composite images (for example, those used in panoramic photography). These could be readily adapted, if not already available for SPM, to construct a composite image according to topography data, and based on this construction show a secondary composite image of the surface potential.

The next step is to analyse this assembled image, reading off the logical states of the cells. Given that the data is already digital and quasi-calibrated, there is clear potential for this to be performed automatically. Obviously, the images gathered are scans of the underside of the memory array, so any readings would have to account for this mirror effect. More troublesome could be any complex security features built into the microcontroller. Given various methods of implementing data encryption, some of these methods could present problems for analysis and would need to be more thoroughly investigated.

To speed up the process and reduce the resulting file size of the composite image (more of an issue than it sounds, *e.g.* hard disk MFM composite images can be terabytes in size), an alternative approach would be to utilise the accuracy of the piezoelectric positioning systems of the AFM to only image the tunnel oxide locations. Scanning just these areas would theoretically reduce the scan time considerably by ignoring the areas between tunnel oxide regions; this would be especially noticeable along the slow scan axis advancement. Various parameters could be manually/automatically determined, *e.g.* marking a reference point at the first tunnel oxide, and the pitch and orientation of the array with relation to this reference. This would decrease the time taken to take each scan, dramatically increasing the throughput of this technique. There is also the possibility that specific data of significance to a forensic investigator could be held in similar locations within the memory arrays of identical microcontrollers. If this were the case, knowledge of the location could dramatically speed up the analysis time required.

It may be possible to increase the robustness of samples after silicon removal has occurred by depositing a layer on top of the circuitry prior to re-encapsulation, forming a new backbone for the thin circuit. This could potentially decrease any interference caused by sample movement/vibration during the initial tapping mode AFM scanning, but would certainly reduce the degree of delamination and thus improve the integrity of samples.

If a backbone were to be safely deposited on top of the circuit to improve the robustness of samples, there is another possible improvement that could be made. Externally connecting and applying a low positive read voltage to individual word lines in a NAND flash array would produce a similar output effect to a normal read operation. Instead of checking for conductivity of the bit line when multiple transistor word lines are pulled high (there is not much hope for channel formation given the lack of silicon substrate), the AFM would measure the increased capacitive effect across the tunnel oxide, dramatically improving the contrast of the scan.

It has already been shown that the SIM chip itself is rather delicate, able to be chipped and damaged with ease, even when handling. However, when encased within the chip module and card body, as a smart card would be during regular operation, it is remarkably robust. The bond wires to the external contact pads are commonly stated as the main cause of (hardware) failure in such devices. With the work outlined in this thesis clearly showing the maximum temperatures that SIM cards could be exposed to before data is destroyed through charge leakage, other questions still remain with regards to the context in which this technique could be applied. One important question that comes to mind is: what is a SIM card's resilience to explosions? With mobile phones commonly being used as detonators for improvised explosive devices (IEDs) and other bombs, will a SIM card microcontroller, presumably containing data pertaining to the history of the SIM card, and of the triggering incoming call, survive such a shock wave?

The most obvious continuation of the quartz grain analysis work is to continue examining both similar and distinctly new grain types, gradually building up a more complete multivariate analysis function. This function will inevitably become more accurate with a greater quantity of data. However, one caveat to bear in mind is that with additional data, it may or may not become more discriminatory. Though distinguishing measures are weighted higher by the func-

tion, new data added for those measures that have a low level of discrimination may decrease its efficacy.

Thus far grains from distinctly different provenance have been examined, with similar distinguishing morphological features and surface textures. However, some areas would undoubtedly contain mixtures of multiple grain types as grain transport occurs. Distinguishing between these grains could prove crucial in determining numerically the ratios encountered in such environments. Such data must be incorporated in any developed database to allow for such samples to be accurately classified.

There are also questions raised during criminal investigations, such as that tackled by Morgan [8] regarding the preservation of quartz grain surface textures following a vehicle fire. The temperatures that quartz grains would be exposed to in such a scenario (810°C) were deemed insufficient to affect the quartz grain surface textures. Results such as these would still need to be verified for AFM analysis given its nanoscale quantitative nature.

At this point, it would be typical to discuss the practical implications of the work. For this thesis, the projects' main objectives have been to research and develop specific applications of the AFM within forensic science. However, there are two other closely linked areas that this work could be applied to. The first is data recovery from damaged solid state storage devices, moving beyond SIM cards and embedded systems and towards dedicated flash memory ICs. Solid state drives (SSDs) have become common PC data storage devices, rapidly replacing magnetic platter hard disk drives (HDDs). With a substantially higher price per unit of storage over traditional magnetic disk drives their adoption was initially low; most commonly filling the roles of high-end laptop drives and PC-enthusiast/workstation dedicated boot drives. They fill both of these niches especially well due to flash memory's negligible seek time and lack of moving parts – resulting in a drive with a greater shock resistance and far lower energy consumption, perfect for mobile computing applications. While they lack the

maximum capacity of their magnetic predecessors, they are catching up – albeit with a hefty price tag for larger capacity SSDs.

Data recovery from magnetic disk drives has been a service available for many years. This often involves simply repairing open connections or replacing faulty components when the drive becomes inoperable, *e.g.* swapping out a faulty PCB or flexible flat (ribbon) cable, replacing a misaligned internal magnetic scan head, or moving the magnetic platter to an identical drive chassis. With a great deal of potentially useful data stored within these devices, forensic data recovery has become a common practice among law enforcement agencies worldwide. This data extraction is typically conducted using dedicated forensically-sound examination suites to clone all existing data, including deleted entries. Sometimes it may be necessary to conduct a deeper microscopic examination of the platter if standard electrical interrogation fails, or to locate deliberately overwritten data pertinent to an investigation. The primary technique for the microscopic analysis and extraction of magnetic media is magnetic force microscopy [9]; while a far lengthier and more costly approach (often taking months), this method can successfully retrieve data from severely damaged drive platters, *e.g.* those that have been drilled into, bent, burned, or cut into several pieces. Even in cases where data has been erased or overwritten, it may still be possible to identify sections of previously stored data due to imperfectly overlapping domains [10] caused by slight aberrations in the rotation speed of the platter and the scan head movement. These resulting 'edge shadows' have been shown to expose the logical state of previously stored bits.

Recovery of data from an SSD usually follows a similar process as for a magnetic drive, with faulty parts replaced or repaired. In certain cases, the flash memory ICs will be desoldered and connected to an identical PCB or a dedicated chip reader. These memory chips are similar to those used in numerous devices, from USB flash drives to mobile phone handsets, from SD memory cards to tablets. The flash memory within these chips typically takes the form of NAND

architecture, and to achieve a higher capacity is most often MLC (usually 2-bit-per-cell, but up to 4-bit-per-cell currently), although some applications warrant the use of lower-capacity SLC for enhanced endurance, increased error margins, and even lower latency. However, as far as the author is aware, there is not as yet a viable method to extract data from damaged (modern) flash memory chips. The successful development of the SIM card data extraction method outlined in this thesis would be an ideal stepping stone towards a subsequent method dealing with flash memory devices. While SIM cards use relatively modern fabrication processes, especially compared to those devices used within astronautical engineering applications, they lag quite a way behind the current CMOS process steps used in NAND flash chips – currently sitting at the 22nm processing step, with most companies producing 25–27nm MLC NAND chips, and some producing even smaller featured variants, *e.g.* Intel's 20nm sync-MLC and Samsung's 21nm Toggle TLC (triple level cell, *i.e.* 3-bit-per-cell MLC).

The second of these closely related applications deals with the classification of Martian dust particles in a similar fashion to the discriminant analysis of quartz sand grains. The goal is to supplement other research areas in identifying characteristics indicative of the historical presence of water on Mars. A couple of feasibility studies have been published in this area [11,12]. Regardless of these studies, it is unknown whether the concept was deemed to be applicable to the micrometre-sized particles examined by the Phoenix lander in 2008, or will be applied to the upcoming 2020 mission, for which the current Curiosity rover is forming the basis. The most obvious issue with such an application is the same as for the forensic analysis of sand grains – the lack of a classifying database of previous measurements to compare and contrast any measurements against.

In conclusion, with the recent criticism levied against the forensic sciences, it is clearer than ever that AFM has great potential within this field for future development. This powerful and versatile technique can be applied in novel applications, to investigate, characterise, and discriminate between sample sets

based on quantitative nanoscale examination. AFM should not be thought of as a ‘one-stop’ technique, but rather as a tool capable of providing independent corroboration. It has the potential to provide calibrated numerical data supplementing existing forensic techniques, and is able to bolster qualitative or semi-quantitative evidentiary claims; but also to disprove ineffective techniques, exposing false assumptions which have plagued judicial systems with controversies in recent years.

Bibliography

[1] C. De Nardi, R. Desplats, P. Perdu, F. Beaudoin, and J. Gauffier, "Oxide charge measurements in EEPROM devices," *Microelectronics and Reliability*, vol. 45, no. 9-11, pp. 1514–1519, 2005.

[2] ——, "EEPROM Failure Analysis Methodology - Can Programmed Charges Be Measured Directly by Electrical Techniques of Scanning Probe Microscopy?" in *31st ISTFA 2005.* San Jose, CA: ASM International, 2005, pp. 256–261.

[3] C. De Nardi, R. Desplats, P. Perdu, J. Gauffier, and C. Guerin, "Descrambling and data reading techniques for flash-EEPROM memories. Application to smart cards," *Microelectronics and Reliability*, vol. 46, no. 9-11, pp. 1569–1574, 2006.

[4] B. J. Jones, "Burnt to memory - data extraction from heat damaged mobile phones," *Public Service Review: Home Office*, no. 15, pp. 68–70, March 2007. [Online]. Available: http://bura.brunel.ac.uk/bitstream/2438/695/1/PSR_HO_Jones_15_2007_68.pdf

[5] B. Jones and A. Kenyon, "Retention of data in heat-damaged SIM cards and potential recovery methods." *Forensic Science International*, vol. 177, no. 1, pp. 42–6, May 2008.

[6] A. D. Purporti Jr. and J. McElroy, "Full-scale house fire experiment for interfire vr - report of test fr 4009," U.S. Department of Commerce, NIST, Gaithersburg, MD, Tech. Rep., 1998. [Online]. Available: http://www.interfire.org/features/fire_experiment.asp

[7] A. Gorbushina, A. Kempe, K. Rodenacker, U. Jütting, W. Altermann, R. Stark, and W. Krumbein, "Quantitative 3-dimensional image analysis

of mineral surface modificationschemical, mechanical and biological," *Geomicrobiology Journal*, vol. 28, no. 2, pp. 172–184, 2011.

[8] R. M. Morgan, M. Little, A. Gibson, L. Hicks, S. Dunkerley, and P. A. Bull, "The preservation of quartz grain surface textures following vehicle fire and their use in forensic enquiry," *Science & Justice*, vol. 48, no. 3, pp. 133–140, Sep. 2008.

[9] D. Rugar, H. Mamin, P. Guethner, S. Lambert, J. Stern, I. McFadyen, and T. Yogi, "Magnetic force microscopy: General principles and application to longitudinal recording media," *Journal of Applied Physics*, vol. 68, no. 3, pp. 1169–1183, 1990.

[10] J.-G. Zhu, Y. Luo, and J. Ding, "Magnetic force microscopy study of edge overwrite characteristics in thin film media," *Magnetics, IEEE Transactions on*, vol. 30, no. 6, pp. 4242–4244, 1994.

[11] A. Kempe, F. Jamitzky, W. Altermann, B. Baisch, T. Markert, and W. Heckl, "Discrimination of aqueous and aeolian paleoenvironments by atomic force microscopya database for the characterization of martian sediments," *Astrobiology*, vol. 4, no. 1, pp. 51–64, 2004.

[12] S. Vijendran, H. Sykulska, and W. Pike, "Afm investigation of martian soil simulants on micromachined si substrates," *Journal of Microscopy*, vol. 227, no. 3, pp. 236–245, 2007.

Appendix A

Lapping/polishing results

Abrasive type	Description	Average particle size (μm)	Initial Ra (nm)			Point-to-point average Ra (nm)	Final Ra (nm)			Point-to-point average Ra (nm)
			1	2	3		1	2	3	
fixed	P1200	15.3	181	190	205	192.0	549	616	464	543.0
fixed	P2500	8.4	46.5	53.2	61.2	53.6	208	219	181	202.7
free-flowing	9μm	9	25.9	26.6	28.8	27.1	70.8	82.5	81.0	78.1
free-flowing	6μm	6	18.1	14.7	12.5	15.1	30.6	34.4	34.9	33.3
free-flowing	1μm	1	7.5	6.8	7.8	7.4	11.3	9.6	10.1	10.3
free-flowing	0.25μm	0.25	4.2	3.1	5.4	4.2	6.7	5.9	6.2	6.3

Appendix B

BOE etching results

Etch time	Step Height (nm) ± 2.5%			Point-to-point average
(min)	1	2	3	step height (nm) ± 2.5%
0	2993	2984	3038	3005.0
10	2472	2470	2523	2488.3
18	2055	2059	2106	2073.3
25	1692	1699	1742	1711.0
30	1436	1438	1483	1452.3
36	1126	1129	1172	1142.3
45	665	665	706	678.7
50	404	405	446	418.3
53	248	250	291	263.0
55	145	147	188	160.0
57	41	44	85	56.7
58	-	-	34	-
59	-	-	-	-

Appendix C

TMAH etching results

Etching solution	Temperature (°C)	Step height (nm) ±1%			Mean ±1%	Data range	Etch rate (nm/min) ±0.2%	Ra (nm)			Mean Ra (nm)
		1	2	3				1	2	3	
	60	551.2	680.7	584.3	605.4	129.5	121.1	3.88	4.12	2.64	3.55
	65	1082.8	1037.4	958.1	1026.1	124.7	205.2	4.73	4.85	3.74	4.44
TMAH 25wt% Aq.	70	1611.3	1404.9	1524.5	1513.6	206.4	302.7	4.78	3.29	4.19	4.09
(hot plate, stirrer)	75	1986.0	2160.4	1997.7	2048.0	174.4	409.6	5.52	4.06	5.81	5.13
	80	2586.1	2403.2	2487.9	2492.4	182.9	498.5	4.41	4.94	6.94	5.43
	85	2713.9	2854.5	2914.8	2827.7	200.9	565.5	5.07	7.46	4.52	5.68
	90	3240.0	3308.8	3065.3	3204.7	243.5	640.9	7.75	5.67	4.53	5.98
	60	479.4	427.3	469.6	458.8	52.1	91.8	2.03	3.03	2.11	2.39
	65	793.4	857.4	680.7	777.2	176.7	155.4	2.38	3.59	2.55	2.84
TMAH 25wt% Aq.	70	1125.3	1031.8	1177.7	1111.6	145.9	222.3	2.54	4.04	3.93	3.50
+ 10% IPA	75	1384.7	1380.1	1256.2	1340.3	128.5	268.1	2.52	4.34	4.03	3.63
(hot plate, stirrer)	80	1756.1	1703.6	1751.9	1737.2	52.5	347.4	5.26	4.97	4.26	4.83
	85	2004.8	2019.5	1904.6	1976.3	114.9	395.3	5.49	4.58	5.29	5.12
	90	2190.5	2320.1	2332.5	2281.0	142.0	456.2	5.26	5.37	4.22	4.95
	60	623.3	609.4	602.7	611.8	20.6	122.4	3.08	4.27	3.32	3.56
	65	956.3	975.0	962.6	964.6	18.7	192.9	4.67	2.67	4.79	4.04
TMAH 25wt% Aq.	70	1384.9	1390.7	1372.1	1382.6	18.6	276.5	4.85	4.26	5.51	4.87
(water bath)	75	1874.0	1854.6	1860.8	1863.1	19.4	372.6	5.86	6.17	4.20	5.41
	80	2226.4	2234.6	2247.4	2236.1	21.0	447.2	6.07	6.13	8.25	6.82
	85	2483.0	2477.1	2495.6	2485.2	18.5	497.0	8.51	8.92	7.19	8.21
	90	2618.1	2604.4	2622.5	2615.0	18.1	523.0	9.96	8.56	8.90	9.14
	60	409.4	413.9	426.6	416.6	17.2	83.3	2.91	2.54	3.45	2.97
	65	708.9	720.2	697.3	708.8	22.9	141.8	2.38	3.25	3.23	2.95
TMAH 25wt% Aq.	70	1008.5	988.9	995.2	997.5	19.6	199.5	3.99	2.83	3.29	3.37
+ 10% IPA	75	1281.6	1298.5	1282.6	1287.6	16.9	257.5	3.05	2.53	3.84	3.14
(water bath)	80	1597.1	1607.9	1586.4	1597.2	21.5	319.4	2.82	3.74	3.78	3.45
	85	1859.8	1881.8	1852.6	1864.7	29.2	372.9	4.74	4.85	4.10	4.56
	90	2155.8	2168.7	2131.1	2151.9	37.6	430.4	5.42	4.76	4.42	4.87

Appendix D

Kelvin probe and baking results

Sample	Type	Before Baking					Baking	After Baking					Rewiring/
		Mean raw work function value (mean 100pts)				Mean surface potential (mV)	temperature	Mean raw work function value (mean 100pts)				Mean surface potential (mV)	reading
		1	2	3	Mean	@ GD=200–205 (mean 100pts)	(°C)	1	2	3	Mean	@ GD=200–205 (mean 100pts)	result
1	G260	-246.6	-241.3	-234.8	-240.9	4659.1	100	-21.4	1.3	23.9	1.3	4901.3	success
2	G260	-445.1	-450.8	-473.9	-456.6	4443.4	200	-824.7	-818.4	-813.7	-818.9	4081.1	success
3	G260	-879.4	-877.9	-882.1	-879.8	4020.2	300	-1543.7	-1519.9	-1501.2	-1521.6	3378.4	success
4	G260	-252.6	-253.0	-243.5	-249.7	4650.3	400	-1864.8	-1856.5	-1815.1	-1845.5	3054.5	unable to read
5	G260	-256.4	-251.6	-257.6	-255.2	4644.8	500	-1022.4	-1015.4	-1009.7	-1015.8	3884.2	unable to read
6	G260	-500.3	-493.7	-491.6	-495.2	4404.8	600	-1172.8	-1169.4	-1165.2	-1169.1	3730.9	unable to read
7	3GP9	-802.7	-816.0	-806.5	-808.4	4091.6	100	-594.6	-597.7	-600.6	-597.6	4302.4	success
8	3GP9	-660.4	-661.9	-665.5	-662.6	4237.4	200	-956.5	-954.4	-952.6	-954.5	3945.5	success
9	3GP9	-1173.3	-1184.1	-1202.7	-1186.7	3713.3	300	-655.5	-656.7	-657.5	-656.6	4243.4	success
10	3GP9	-848.8	-855.0	-871.4	-858.4	4041.6	400	-656.8	-664.1	-665.4	-662.1	4237.9	success
11	3GP9	-954.5	-982.5	-955.9	-964.3	3935.7	500	-1262.8	-1258.2	-1254.1	-1258.4	3641.6	success
12	3GP9	-898.4	-898.6	-899.7	-898.9	4001.1	600	-990.8	-990.4	-990.2	-990.5	3909.5	unable to read
13	3GP9	-880.0	-882.3	-898.7	-887.0	4013.0	400	-1207.4	-1196.3	-1220.8	-1208.2	3691.8	success
14	L254	-725.5	-716.0	-727.5	-723.0	4177.0	100	-1349.3	-1337.5	-1333.1	-1340.0	3560.0	unable to read
15	L254	-449.8	-492.1	-503.2	-481.7	4418.3	200	1370.4	1399.2	1383.8	1384.5	6284.5	damage-tweezers
16	L254	-368.8	-375.0	-367.1	-370.3	4529.7	300	-649.0	-647.4	-649.8	-648.7	4251.3	unable to read
17	L254	-420.1	-423.6	-438.2	-427.3	4472.7	400	-648.4	-644.9	-642.9	-645.4	4254.6	unable to read
18	L254	-1083.4	-1061.0	-1047.3	-1063.9	3836.1	500	-985.2	-984.6	-984.1	-984.6	3915.4	damage-wire bonder
19	L254	-326.7	-320.6	-316.6	-321.3	4578.7	600	-764.3	-761.5	-759.6	-761.8	4138.2	unable to read
20	L254	-234.2	-231.0	-219.4	-228.2	4671.8	100	-576.1	-579.6	-582.6	-579.4	4320.6	unable to read
21	L254	-287.9	-306.4	-300.9	-298.4	4601.6	200	-86.0	-89.8	-83.6	-86.5	4813.5	unable to read
22	L254	-851.3	-859.9	-841.2	-850.8	4049.2	300	-606.3	-608.5	-611.0	-608.6	4291.4	unable to read
23	L254	-	-	-	-	-	-	-	-	-	-	-	damage-decap
24	assorted	-260.6	-235.2	-248.2	-248.0	4652.0	500	-663.7	-667.3	-672.9	-668.0	4232.0	success
25	assorted	-1002.3	-998.1	-990.0	-996.8	3903.2	600	-423.1	-429.9	-437.8	-430.3	4469.7	damage-wire bonder
26	assorted	-526.5	-525.2	-524.2	-525.3	4374.7	100	-100.0	-105.8	-112.2	-106.0	4794.0	success
27	assorted	-	-	-	-	-	-	-	-	-	-	-	damage-decap
28	assorted	-1743.3	-1806.9	-1806.9	-1785.7	3114.3	200	-780.3	-774.1	-770.7	-775.0	4125.0	success
29	assorted	-1947.7	-1949.0	-1974.0	-1956.9	2943.1	300	325.5	323.3	321.5	323.4	5223.4	success
30	assorted	-960.1	-962.1	-971.6	-964.6	3935.4	400	-822.2	-825.1	-828.8	-825.4	4074.6	unable to read
31	assorted	-887.4	-883.2	-864.9	-878.5	4021.5	500	-1299.4	-1293.1	-1290.3	-1294.3	3605.7	unable to read